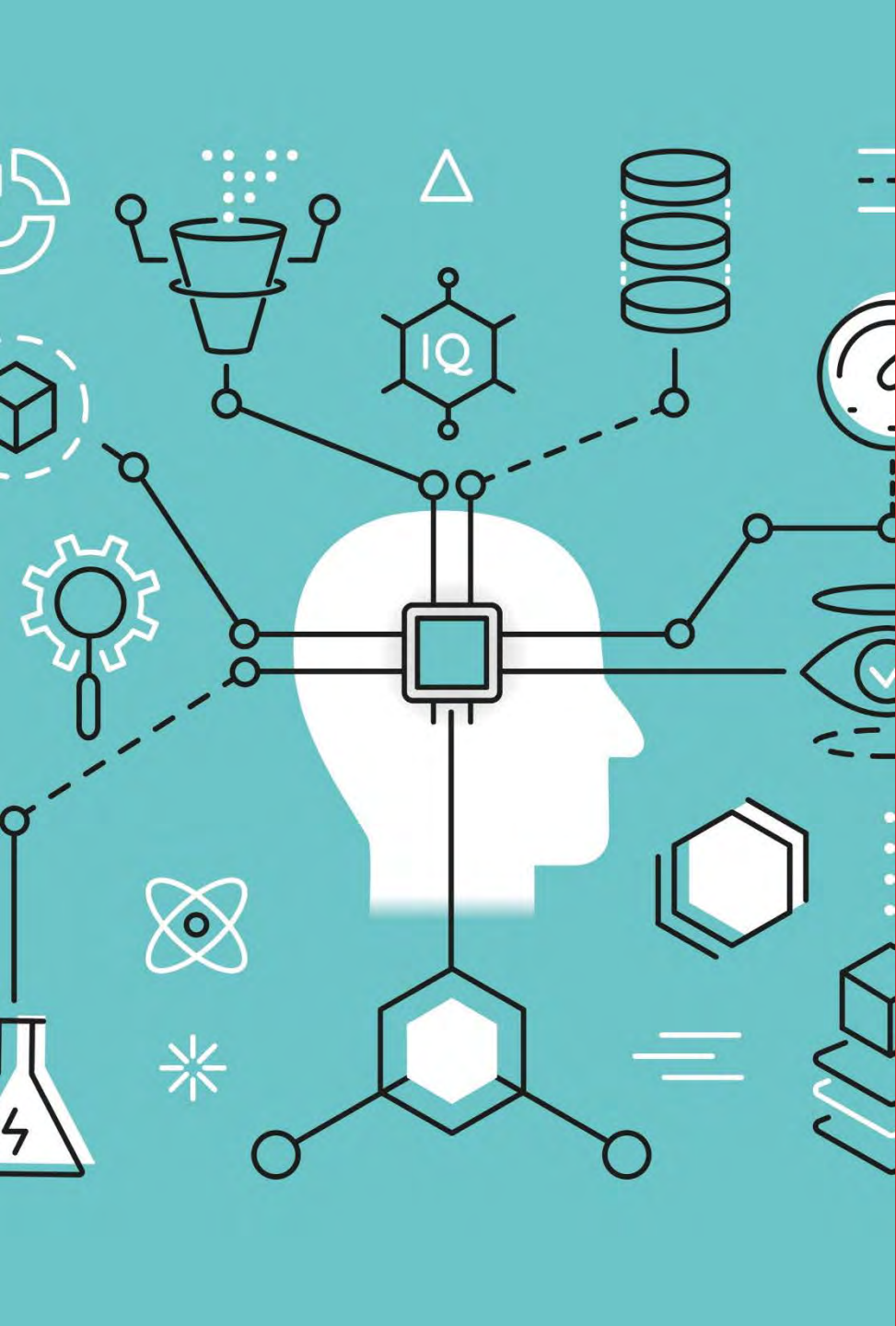


# **2018 Predictive Analytics Symposium**

## **Session 21: Dangers of Overfitting; Myths and Facts of Predictive Analytics (PA)**

[SOA Antitrust Compliance Guidelines](#)

[SOA Presentation Disclaimer](#)



**RGA**

# Dangers of Overfitting in Predictive Analytics

Rosmery Cruz

# Outline

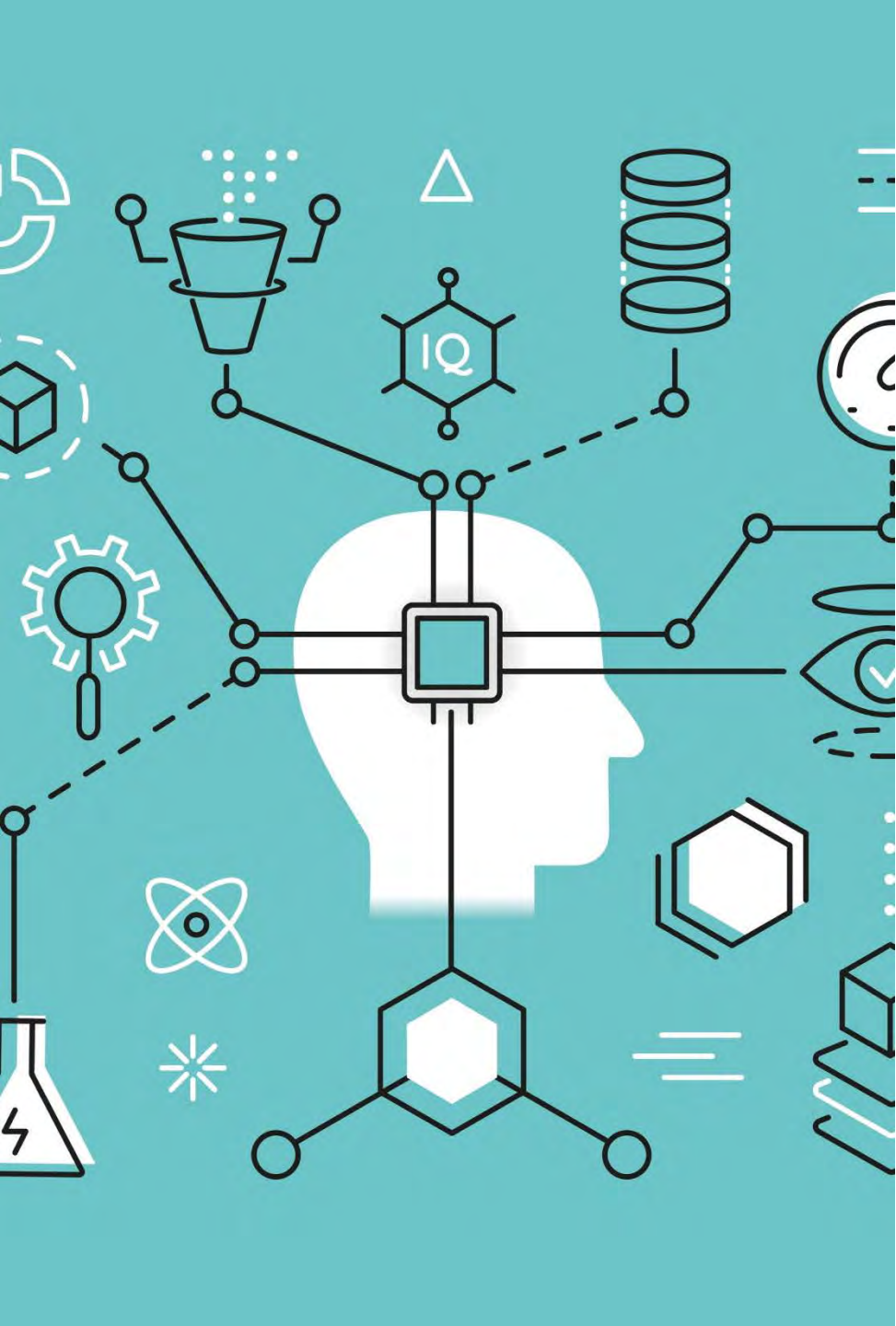
## 1 Motivation

## 2 Overfitting

- Definition

- When Does Overfitting Occur?

- How Do You Prevent Overfitting?



**RGA**

## Motivation

# Power Posing

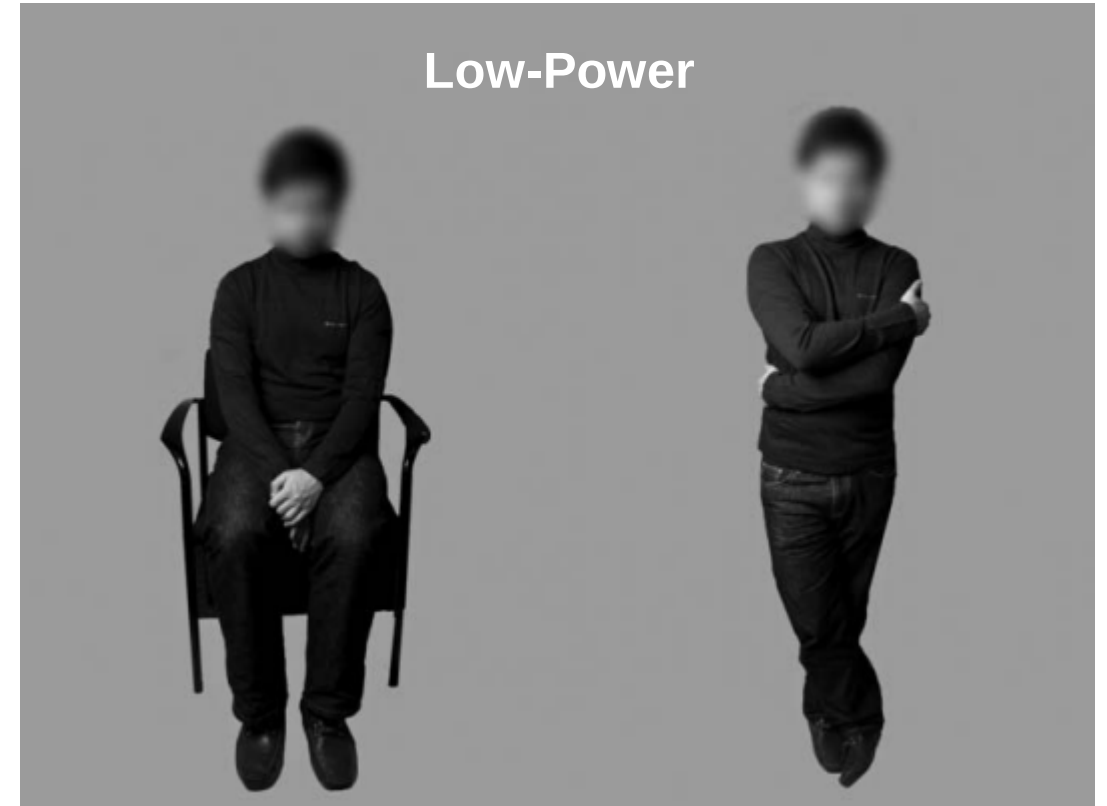
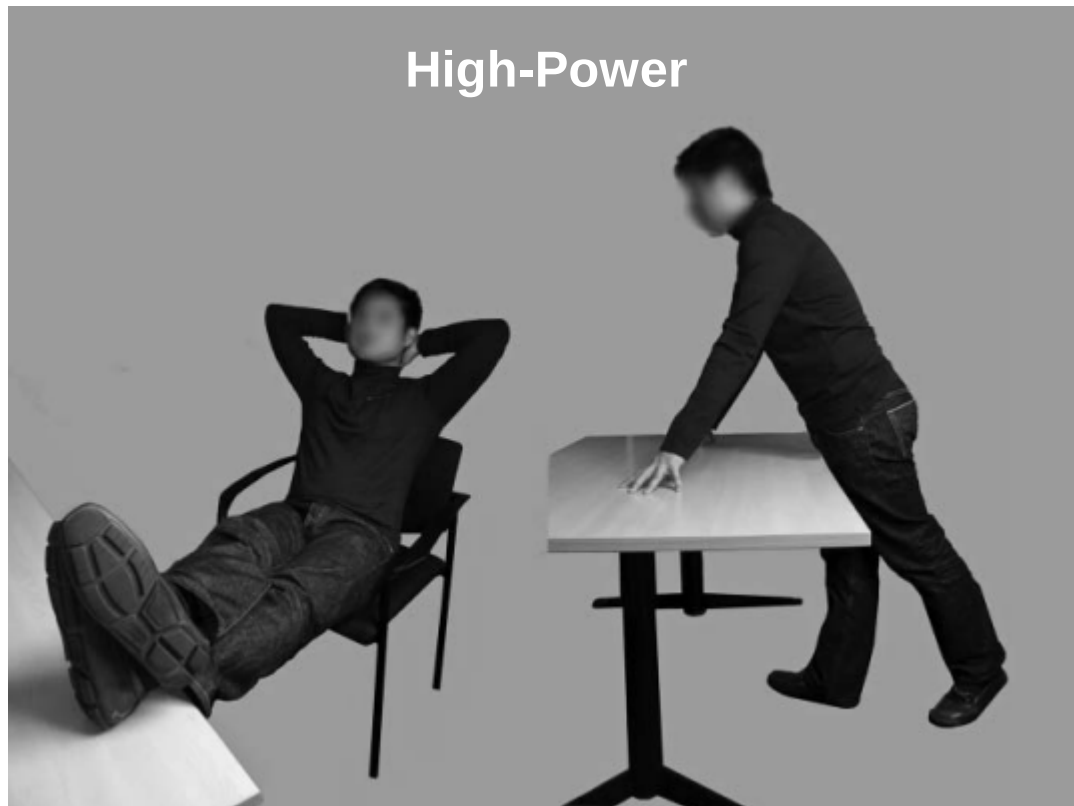
Most popular **TED** talk  
of all time



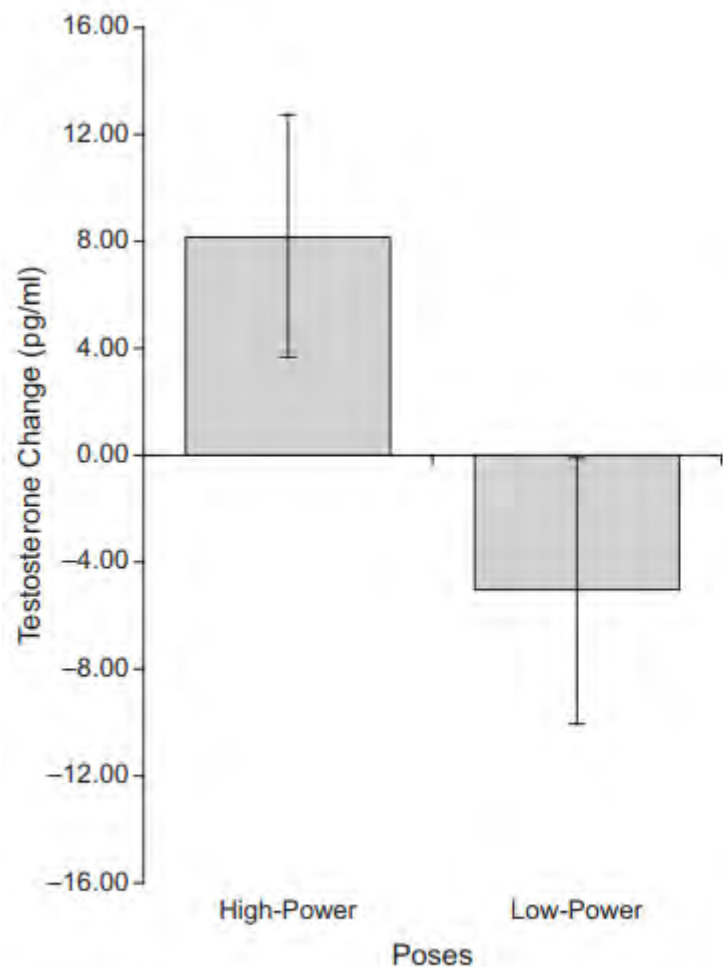
# Power Posing

Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance

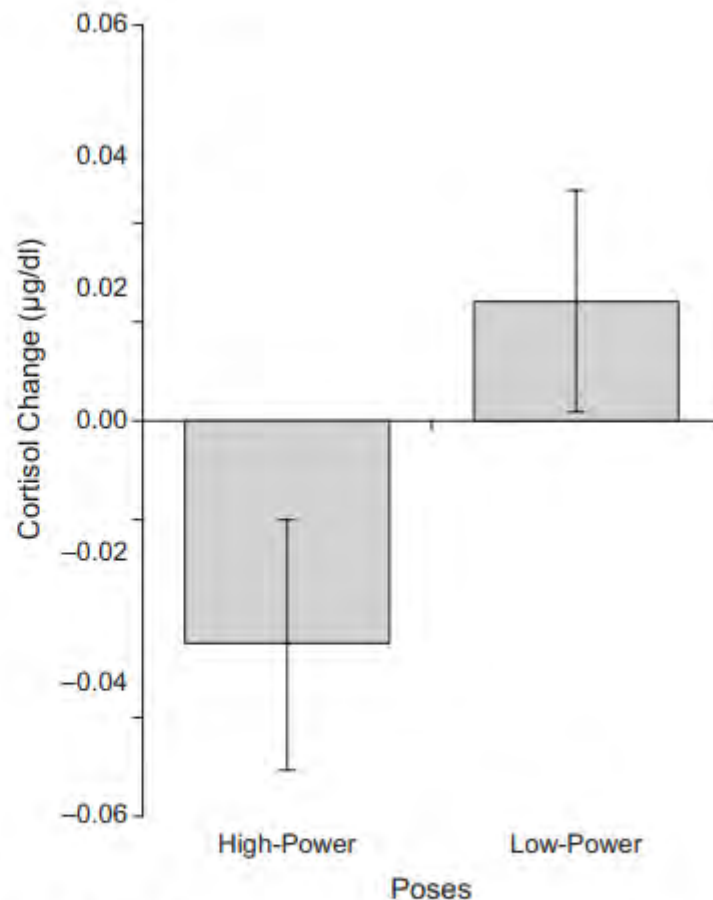
Authors: Dana R. Carney , Amy J.C. Cuddy, and Andy J. Yap



# Power Posing



**Fig. 3.** Mean changes in the dominance hormone testosterone following high-power and low-power poses. Changes are depicted as difference scores (Time 2 – Time 1). Error bars represent standard errors of the mean.



**Fig. 4.** Mean changes in the stress hormone cortisol following high-power and low-power poses. Changes are depicted as difference scores (Time 2 – Time 1). Error bars represent standard errors of the mean.

## Findings

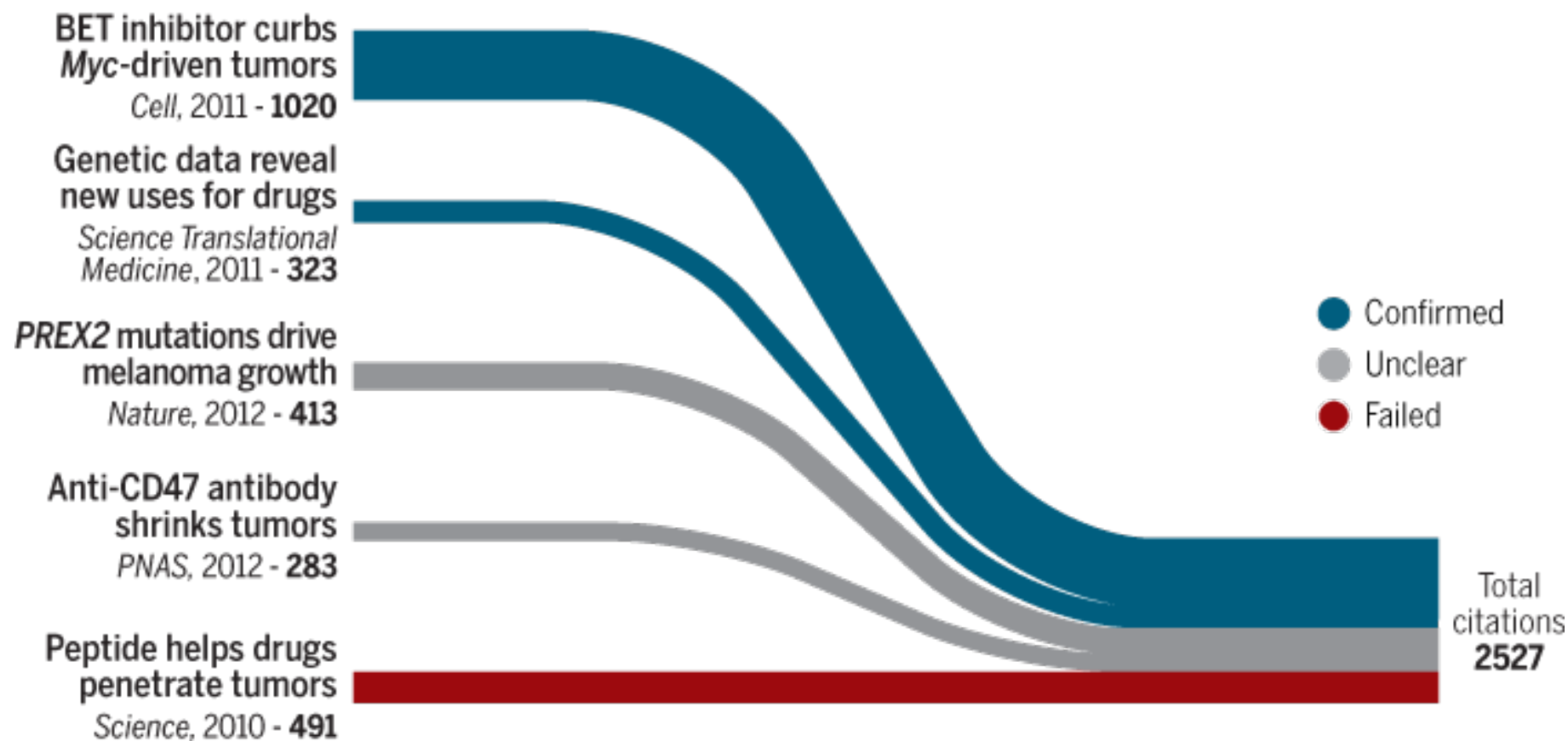
Increased testosterone levels/lower cortisol levels among high-power posers

High-power posers were more likely than low-power poses to take gambling risk

Eleven new  
studies  
suggest  
'Power Poses'  
don't work



# The Replication Crisis



**Power pose is not unique.**

In 2015, two thirds of psychology studies failed replication tests.

Cancer studies have faced similar problems with non-replicable findings.

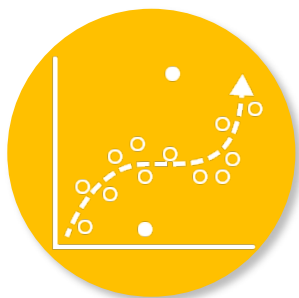
# Motivation: Building Predictive Models



We are asked to build predictive models

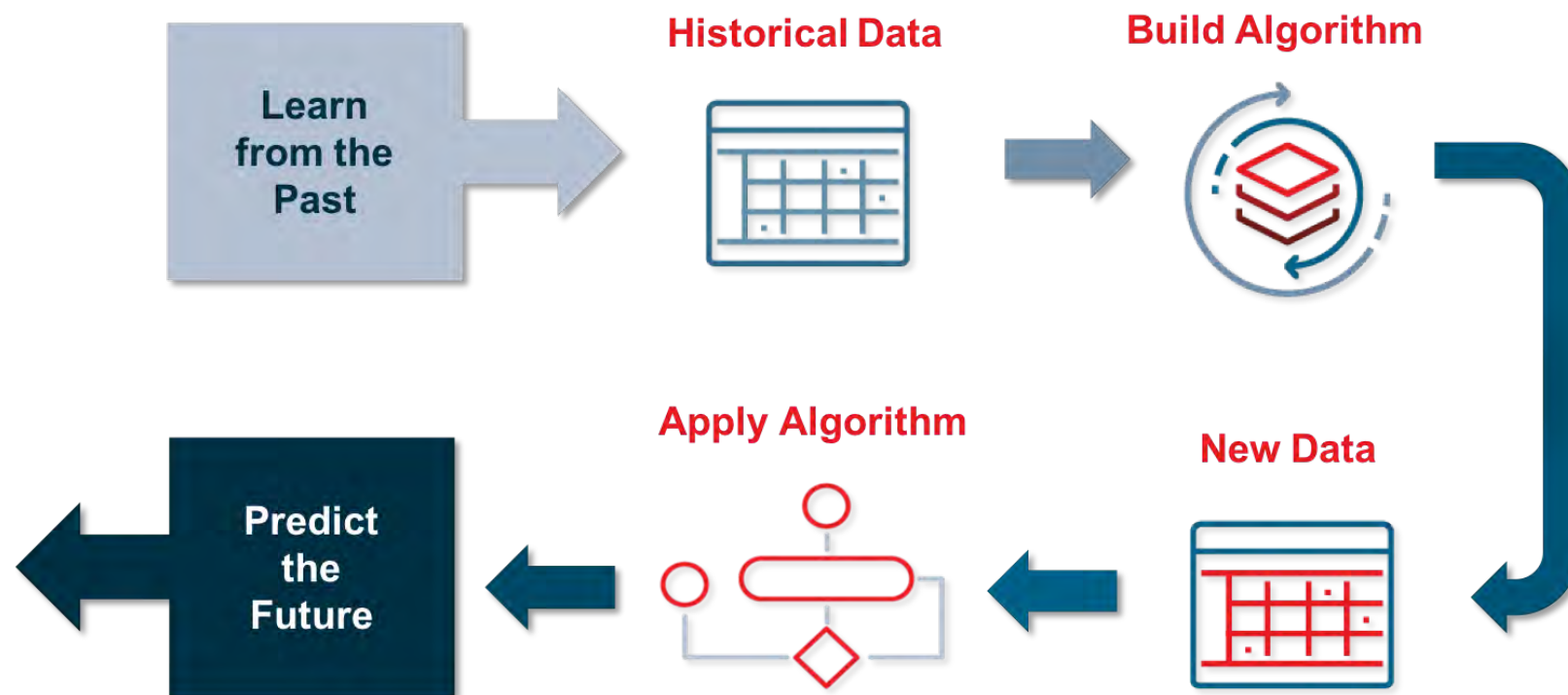


We are given a fixed set of data

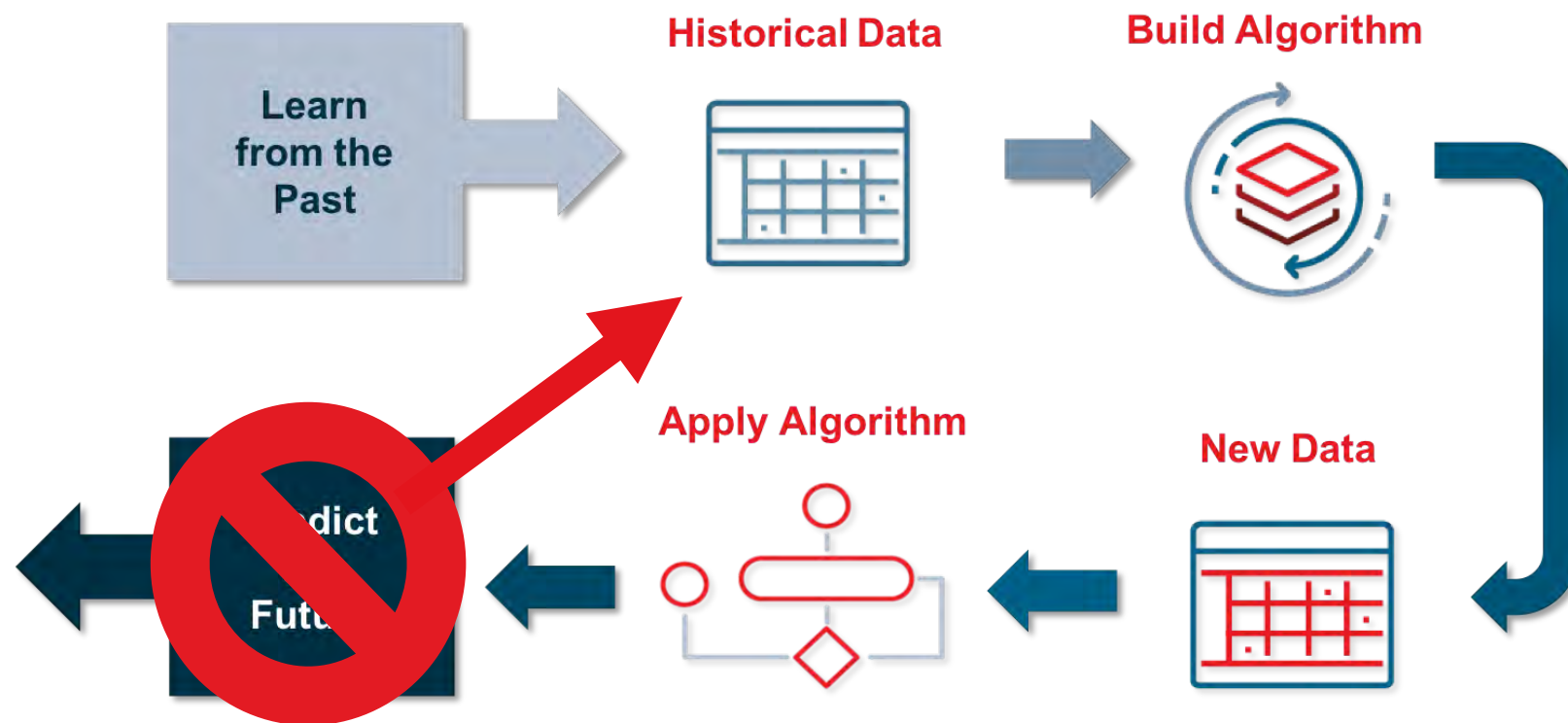


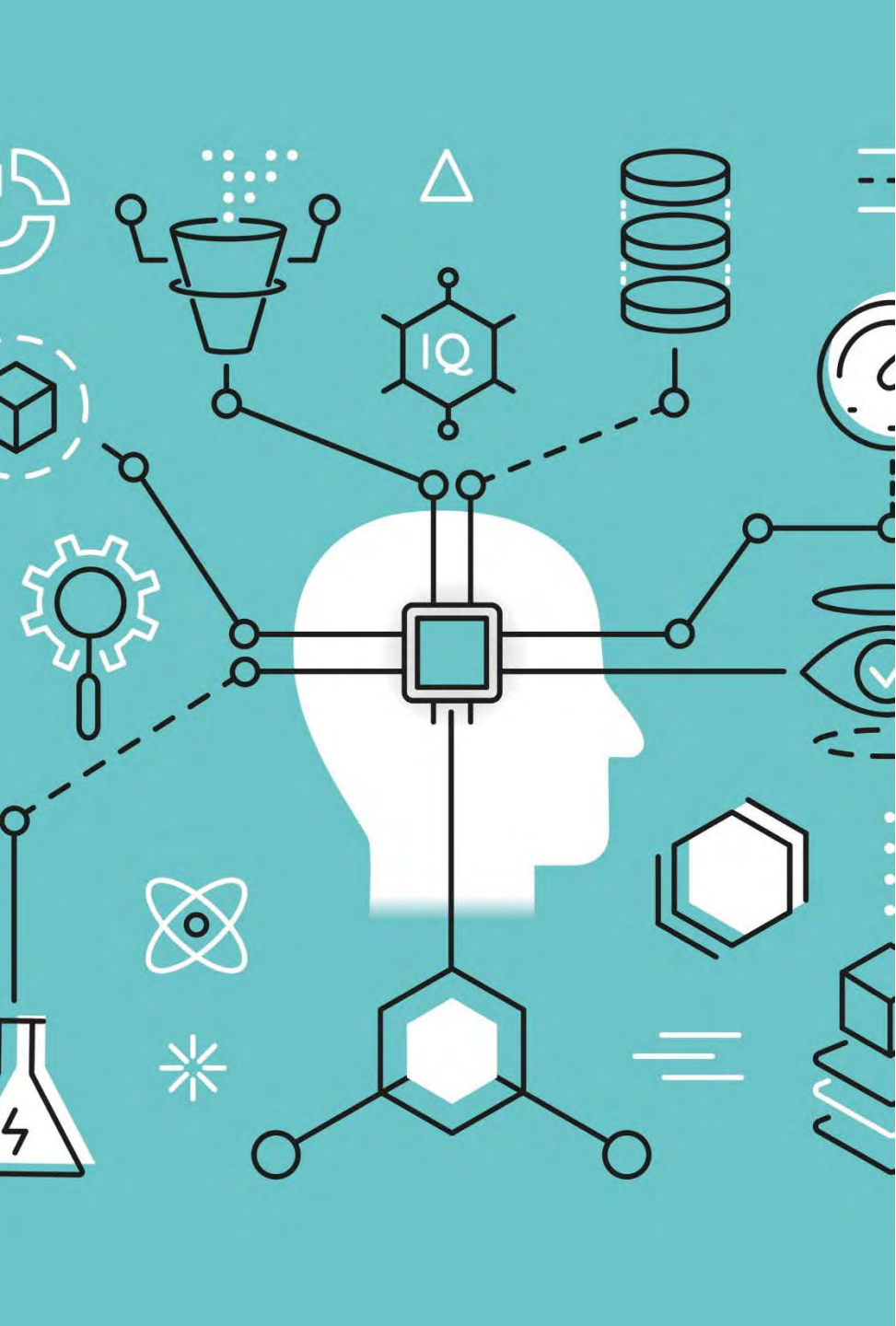
**PROBLEM:** How do we know our model will predict new data reasonably well?

# Motivation: Building Predictive Models



# Motivation: Building Predictive Models





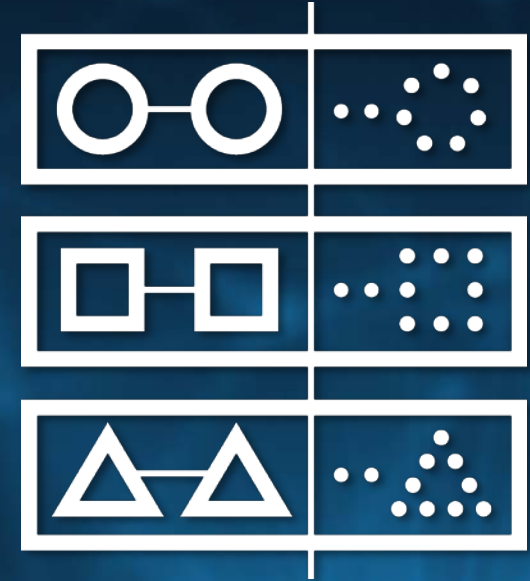
**RGA**

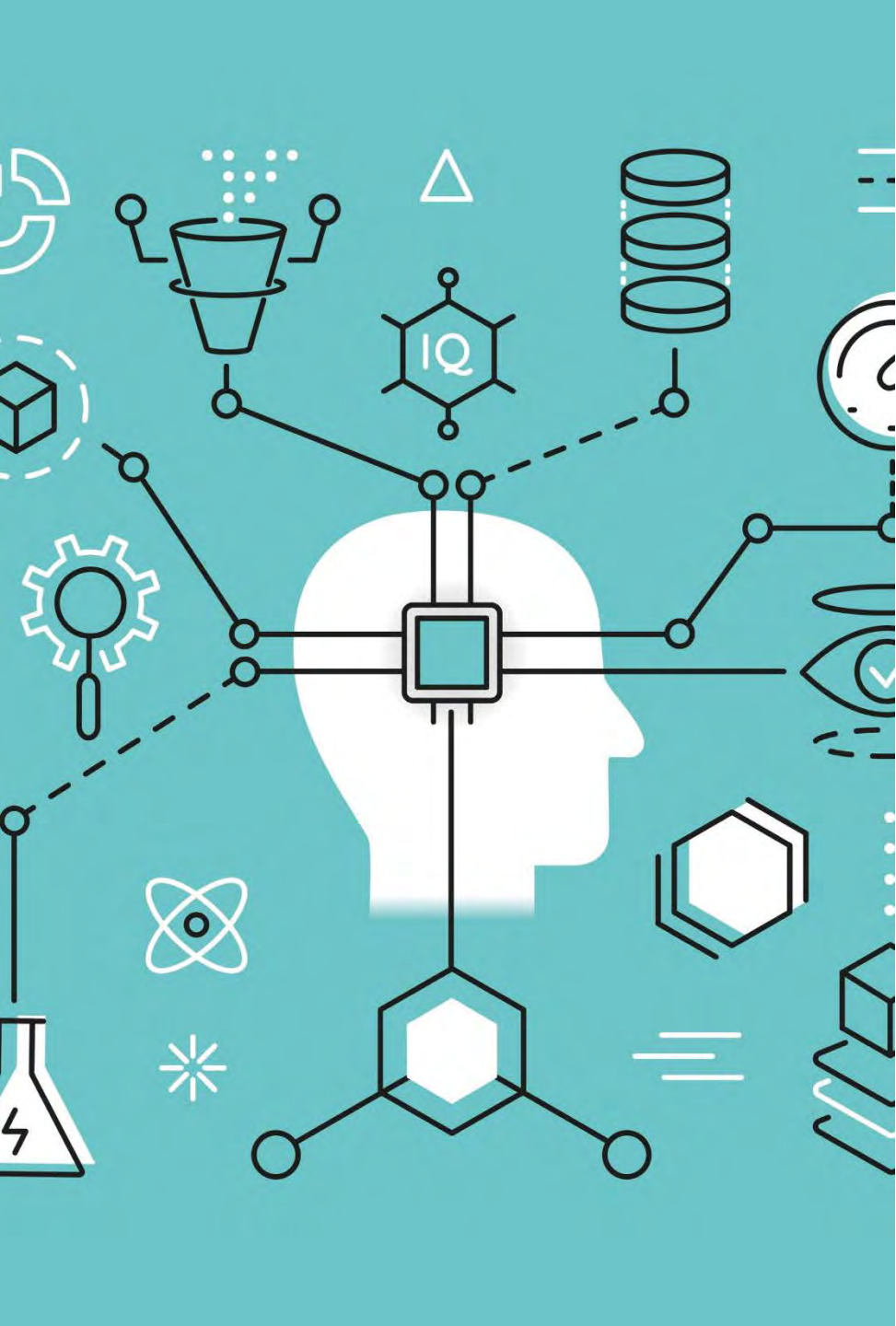
## Overfitting: A Definition

# Overfitting: Definition

“The problem of capitalizing on the idiosyncratic characteristics of the sample at hand, also known as *overfitting*, in regression-type models.

Overfitting yields overly optimistic model results: “findings” that appear in an overfitted model don’t really exist in the population and hence will not replicate.” (Babyak, 2004)





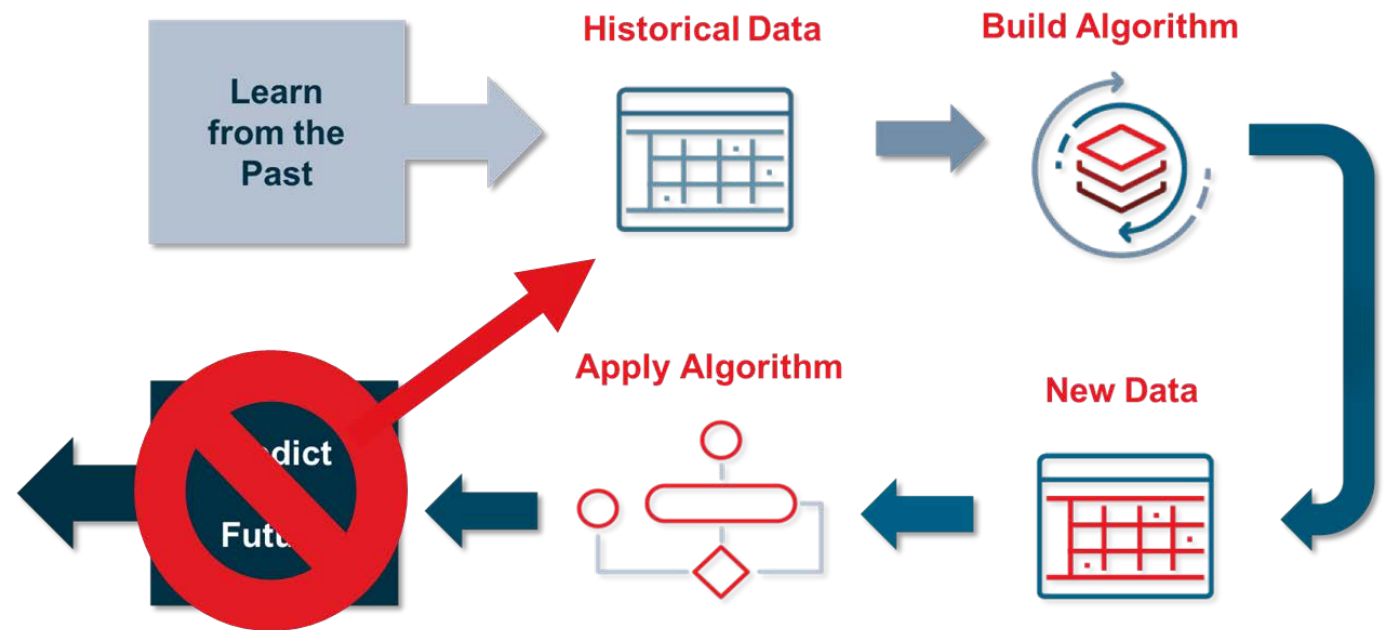
**RGA**

## When Does Overfitting Occur?

# When Does Overfitting Occur?

Generally, overfitting occurs due to analyst oversight in two key areas:

- **Researcher degrees of freedom** (also known as procedural overfitting, data dredging, p-hacking, etc.)
- **Asking too much from the data** (model complexity)

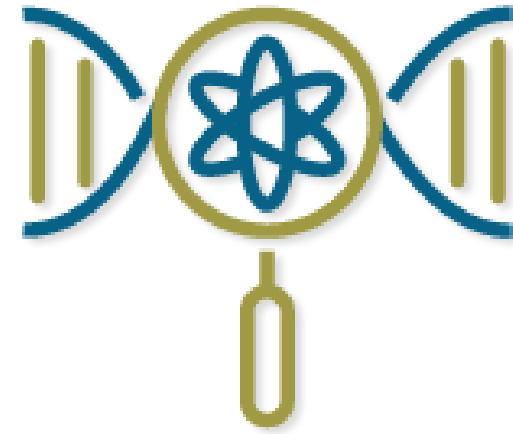


# When Does Overfitting Occur?

Researcher Degrees of Freedom

## Example:

Dataset of 1000 individuals for a weight-loss biomarker study with three time points



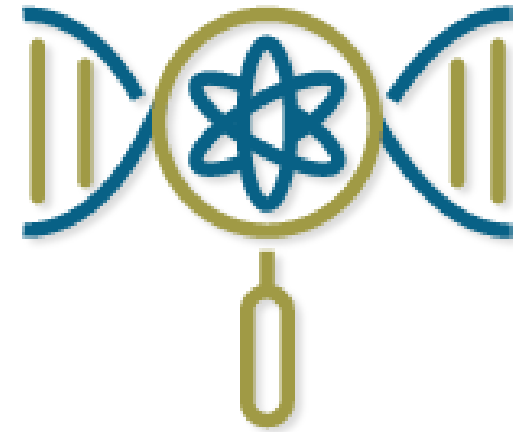
# When Does Overfitting Occur?

## Researcher Degrees of Freedom

Bob performs some simple data exploration.

**He first uses data visualization** to investigate the average activity of all the genes across all the individuals at each of the time points, and **observes that there is very little difference** between time 1 and 2 and there is a **large jump between** time 2 and 3 in the average activity.

So **he decides** to focus on these later two time points.



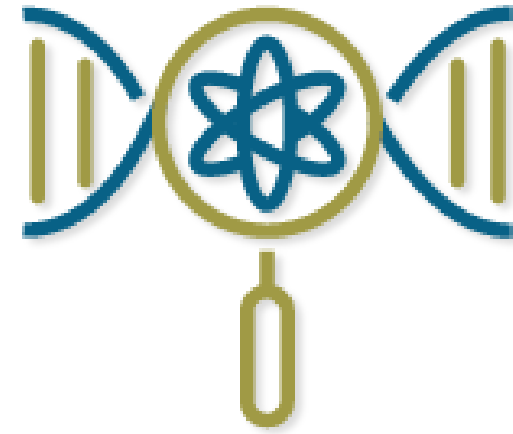
Example and text from Russo and Zou 2016 How much does your data exploration overfit? Controlling bias via information usage.

The word “expression” was replaced with “activity” for simplification.

# When Does Overfitting Occur?

Researcher Degrees of Freedom

Next, he realizes that **half of the genes** always have low activity values and decides to simply filter them out.



Example and text from Russo and Zou 2016 How much does your data exploration overfit? Controlling bias via information usage.

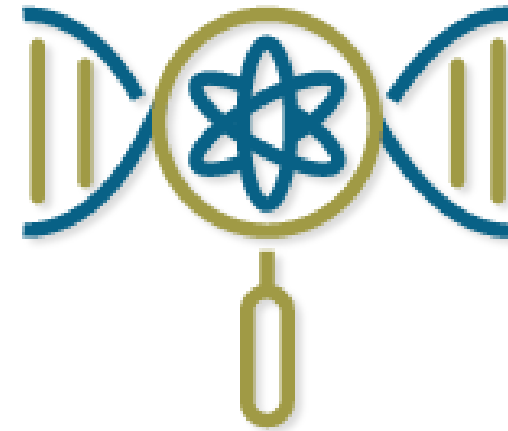
The word “expression” was replaced with “activity” for simplification.

# When Does Overfitting Occur?

## Researcher Degrees of Freedom

Finally, he **computes the correlations** between the activity of the 1000 post-filtered genes and the weight change between time 2 and 3.

He selects the gene with the largest correlation and reports its value.



Example and text from Russo and Zou 2016 How much does your data exploration overfit? Controlling bias via information usage.

The word “expression” was replaced with “activity” for simplification.

# When Does Overfitting Occur?

Researcher Degrees of Freedom



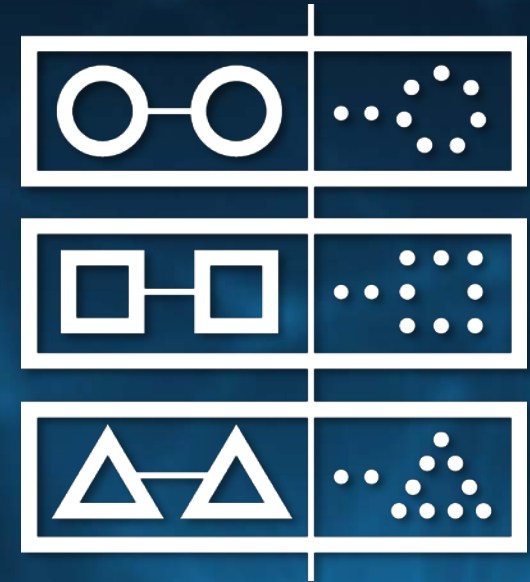
What did  
Bob  
do  
wrong?

# When Does Overfitting Occur?

The culprit is a construct we refer to as **researcher degrees of freedom**. In the course of collecting and analyzing data, researchers have many decisions to make:

- Should more data be collected?
- Should some observations be excluded?
- Which conditions should be combined and which ones compared?
- Which control variables should be considered?
- Should specific measures be combined or transformed or both?

-Simmons, Nelson, and Simonsohn, 2011



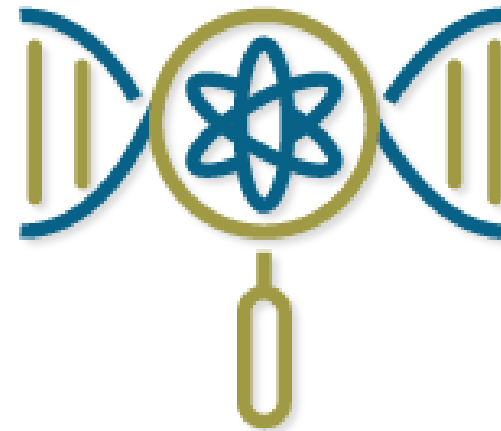
# When Does Overfitting Occur?

## Researcher Degrees of Freedom

Bob performs some simple data exploration.

He first uses **data visualization** to investigate the average activity of all the genes across all the individuals at each of the time points, and **observes that there is very little difference** between time 1 and 2 and there is a **large jump between** time 2 and 3 in the average activity.

So **he decides** to focus on these later two time points.



Research design decisions shouldn't be contingent on observed results. Use previous experience or knowledge to guide analysis choices.

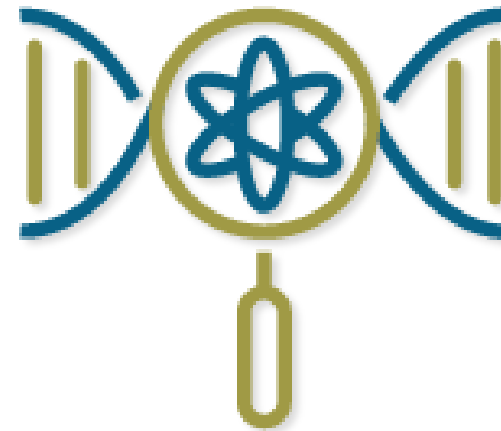
Example and text from Russo and Zou 2016 How much does your data exploration overfit? Controlling bias via information usage.

The word "expression" was replaced with "activity" for simplification.

# When Does Overfitting Occur?

## Researcher Degrees of Freedom

Next, he realizes that **half of the genes** always have low activity values and decides to simply filter them out.



What are “low” activity values? These decisions may be arbitrary. If they’re determined by this dataset, it may not generalize.

Example and text from Russo and Zou 2016 *How much does your data exploration overfit? Controlling bias via information usage.*

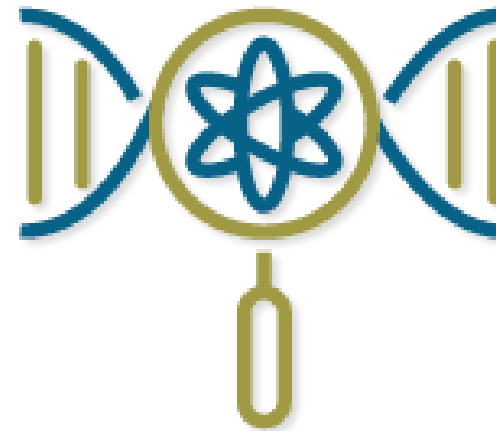
The word “expression” was replaced with “activity” for simplification.

# When Does Overfitting Occur?

## Researcher Degrees of Freedom

Finally, he **computes the correlations** between the activity of the 1000 post-filtered genes and the weight change between time 2 and 3.

He selects the gene with the largest correlation and reports its value.



The resulting “largest” correlation is built upon the series of analysis choices made before it. Again, may not generalize.

Example and text from Russo and Zou 2016 *How much does your data exploration overfit? Controlling bias via information usage.*

The word “expression” was replaced with “activity” for simplification.

# When Does Overfitting Occur?

Researcher Degrees of Freedom

1

Make research design decisions before analyzing the data

2

Where applicable, use subject matter knowledge to inform data aggregation (i.e., age groups)

3

Limit the exclusion of data

4

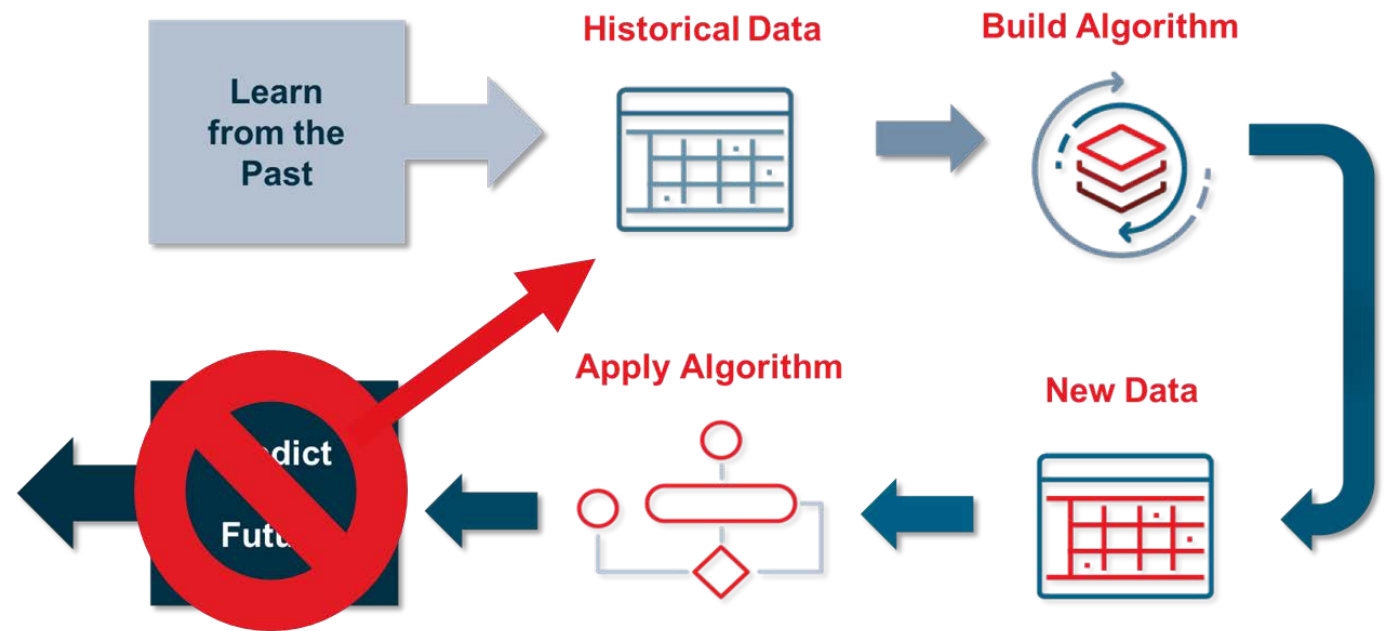
Validate your results (discussed later in the presentation)

Strategies to  
minimize  
researcher  
degrees of  
freedom

# When Does Overfitting Occur?

Generally, overfitting occurs due to analyst oversight in two key areas:

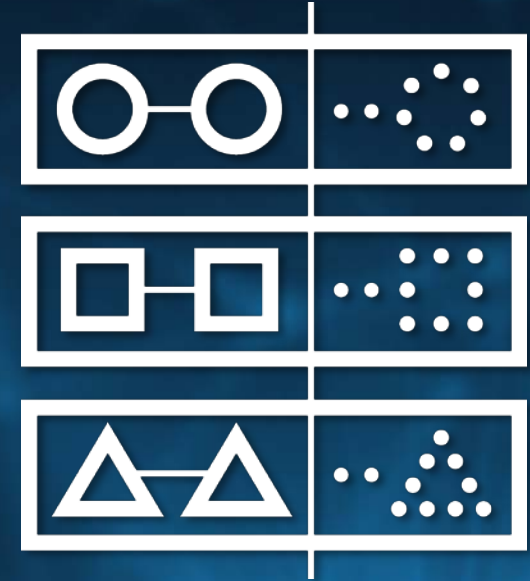
- Researcher degrees of freedom (also known as procedural overfitting, data dredging, p-hacking, etc.)
- Asking too much from the data (model complexity)



# When Does Overfitting Occur?

“Given a certain number of observations in a data set, there is an upper limit to the complexity of the model that can be derived with any acceptable degree of uncertainty.”  
(Babiyak, 2004)

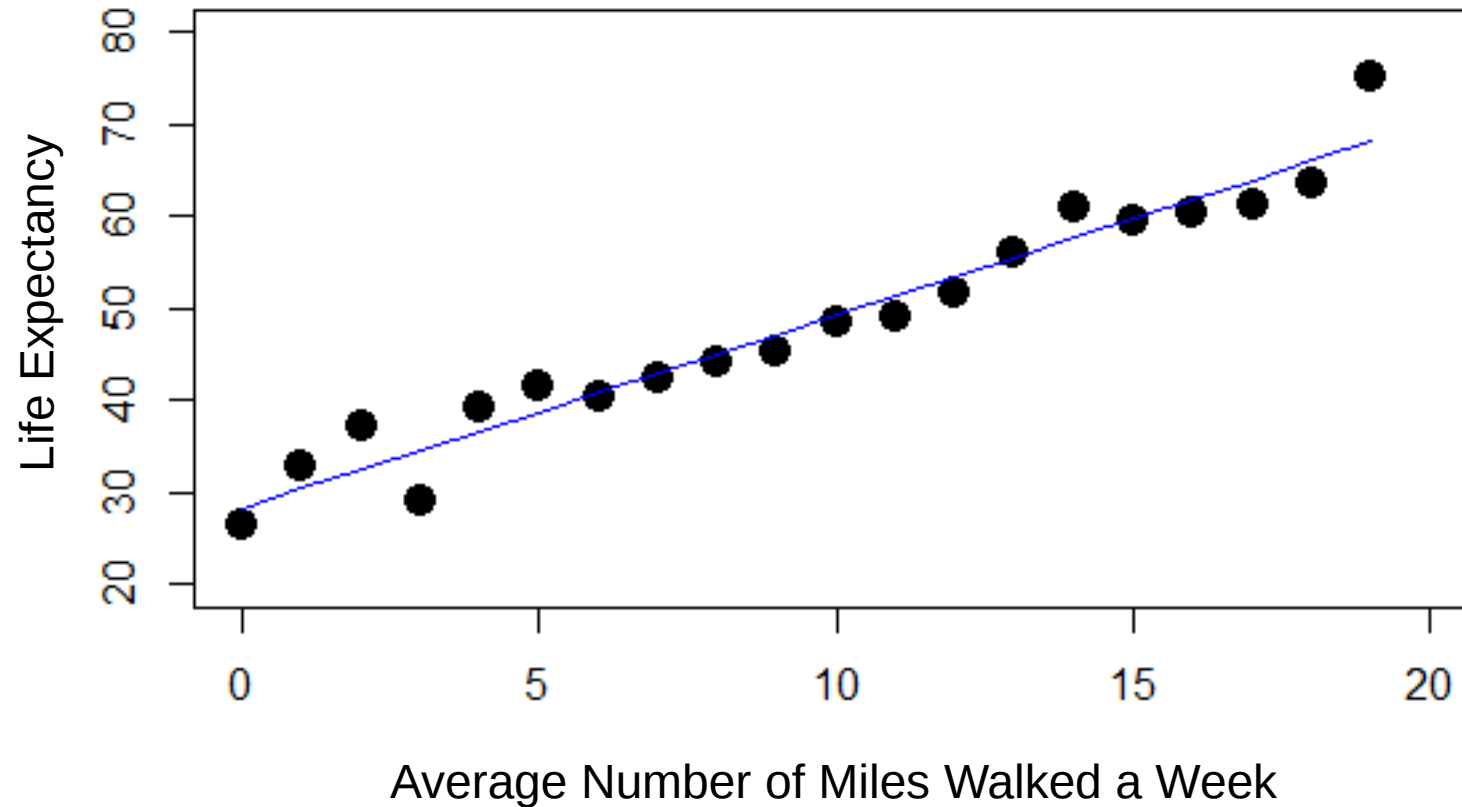
...asking too much from the data



# When Does Overfitting Occur?

Sample Size & Model Complexity

Example: Simulated Data  
N: 20

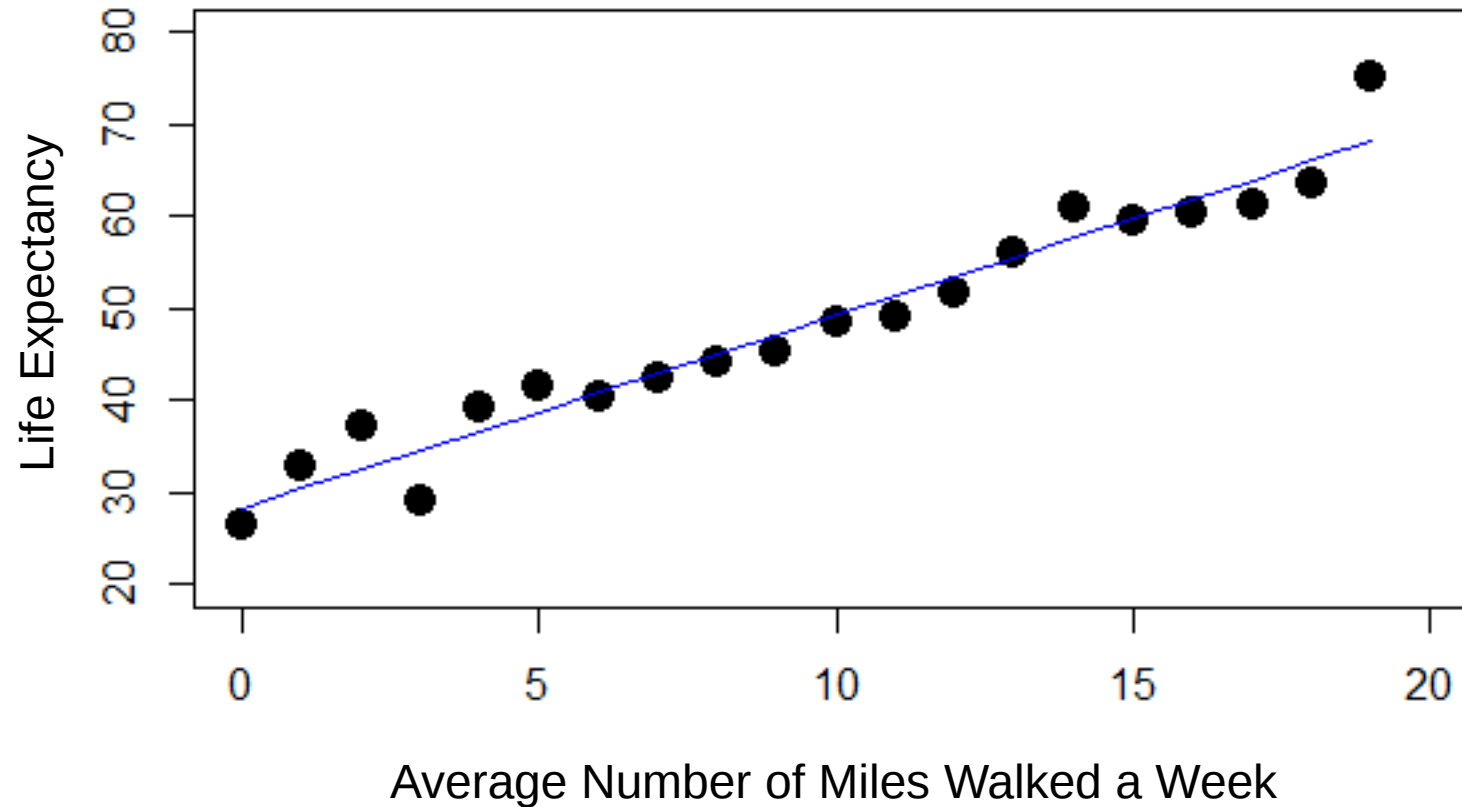


Let's build  
a model

# When Does Overfitting Occur?

Sample Size & Model Complexity

Example: Simulated Data  
N: 20



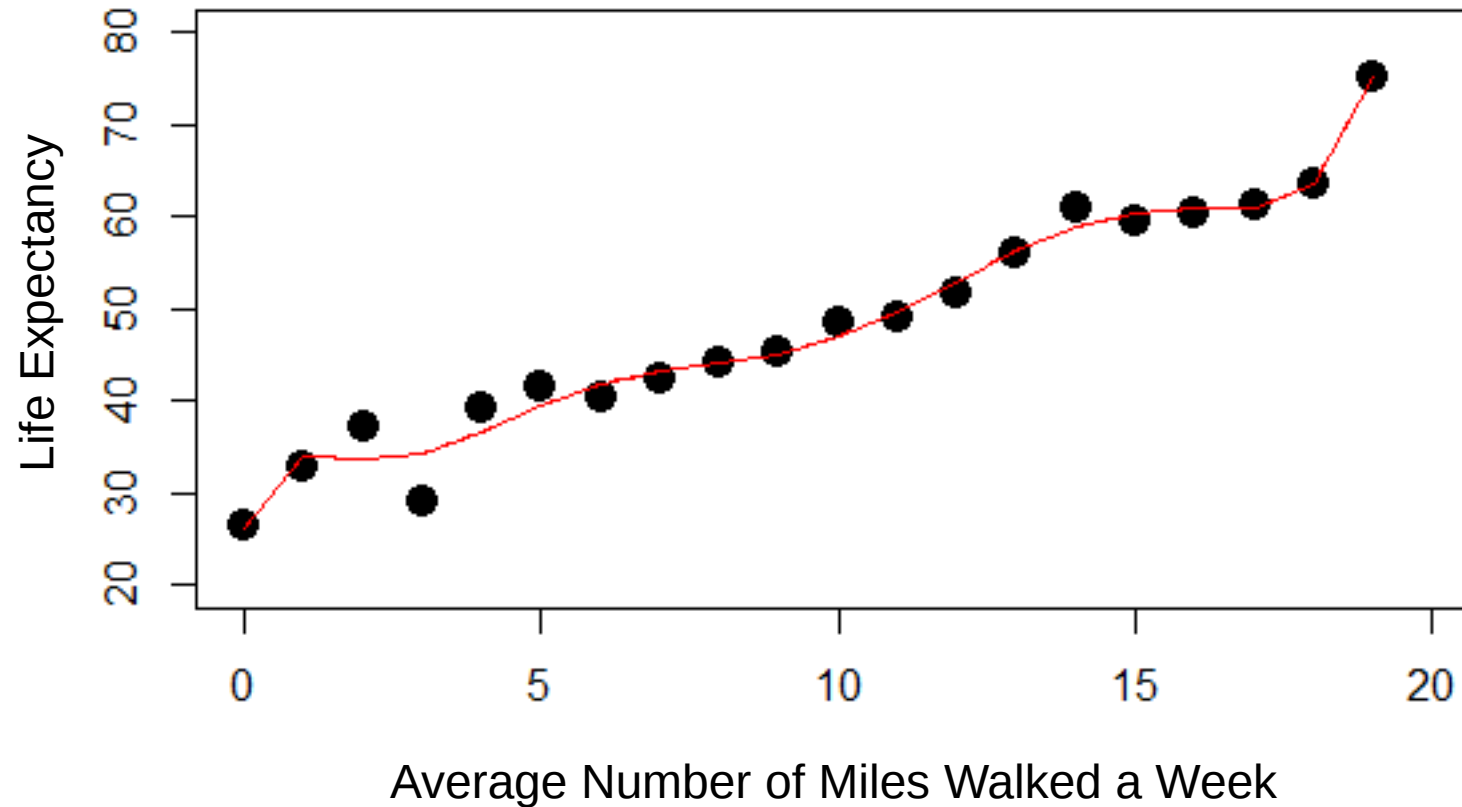
Simple Model  
 $Y = \beta_0 + \beta_1 X$

Simple Model Error  
MSE: 8.45

# When Does Overfitting Occur?

Sample Size & Model Complexity

Example: Simulated Data  
N: 20



Complex Model  
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_8 X^8$$

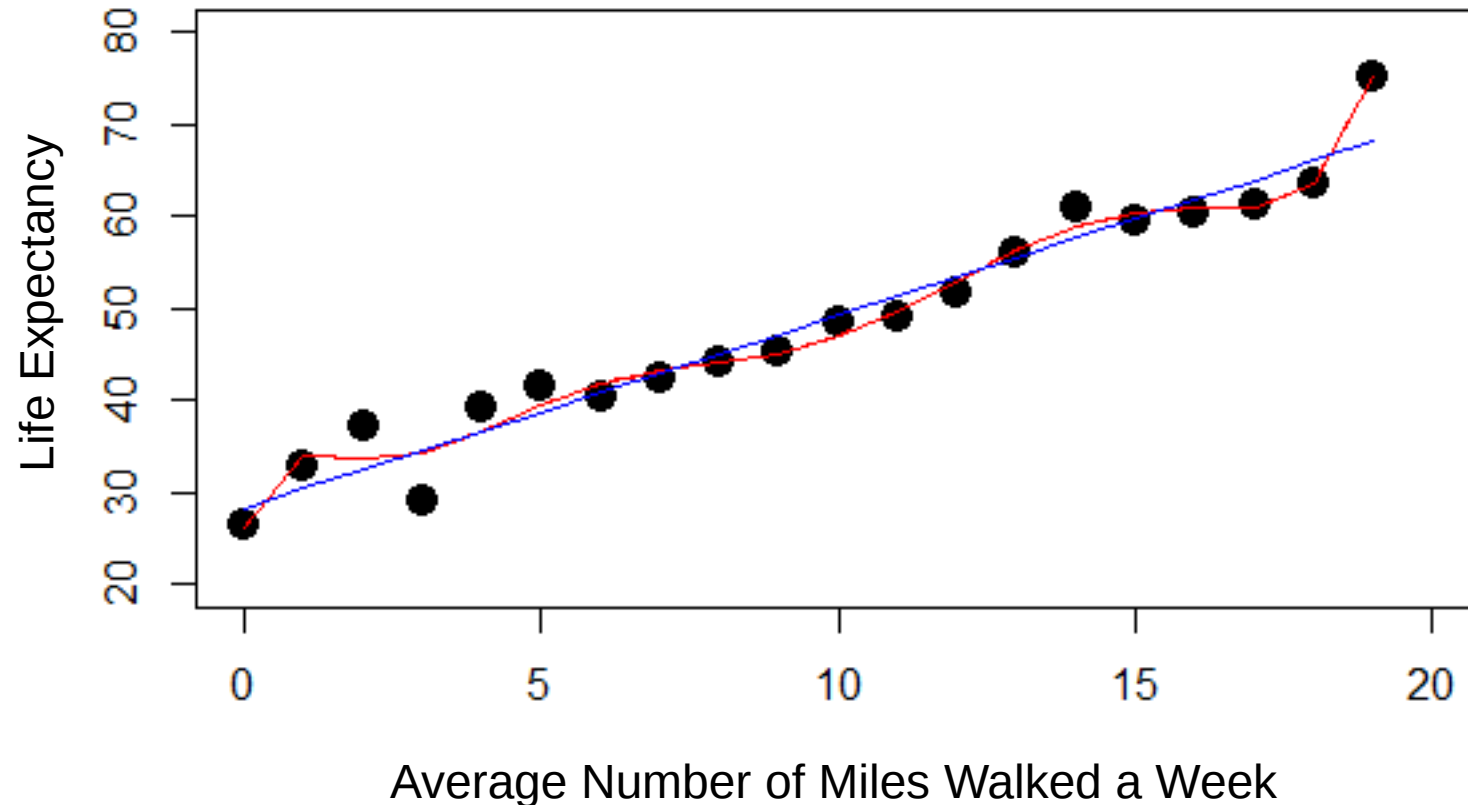
Complex Model Error  
MSE: 3.27



# When Does Overfitting Occur?

Sample Size & Model Complexity

Example: Simulated Data  
N: 20



Which model fits this dataset better?

Simple Model

$$Y = \beta_0 + \beta_1 X$$

Complex Model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_8 X^8$$

Simple Model Error

MSE: 8.45

Complex Model Error

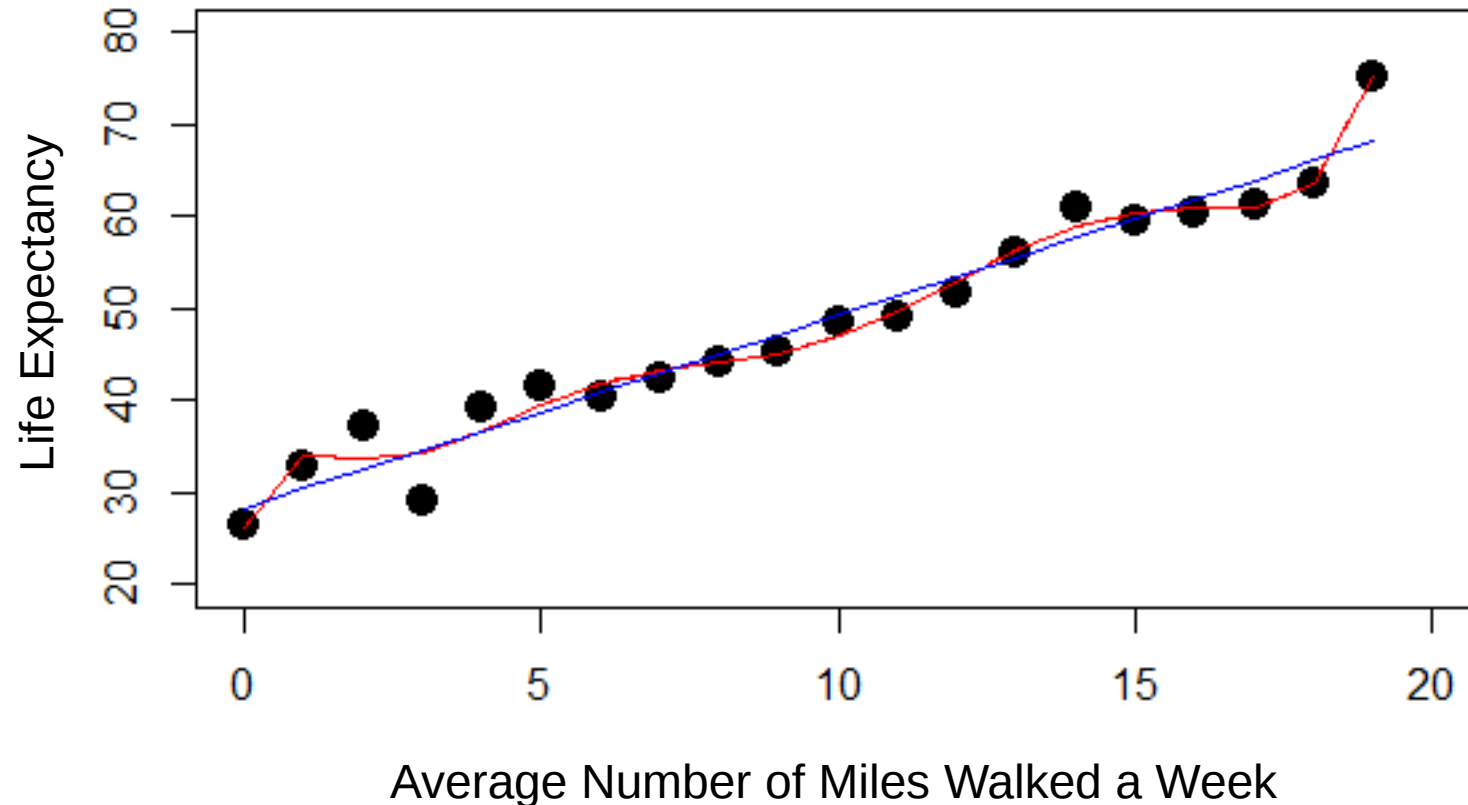
MSE: 3.27



# When Does Overfitting Occur?

Sample Size & Model Complexity

Example: Simulated Data  
N: 20



Which model fits this dataset better?

Simple Model

$$Y = \beta_0 + \beta_1 X$$

Complex Model

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_8 X^8$$

Simple Model Error

MSE: 8.45



Complex Model Error

MSE: 3.27

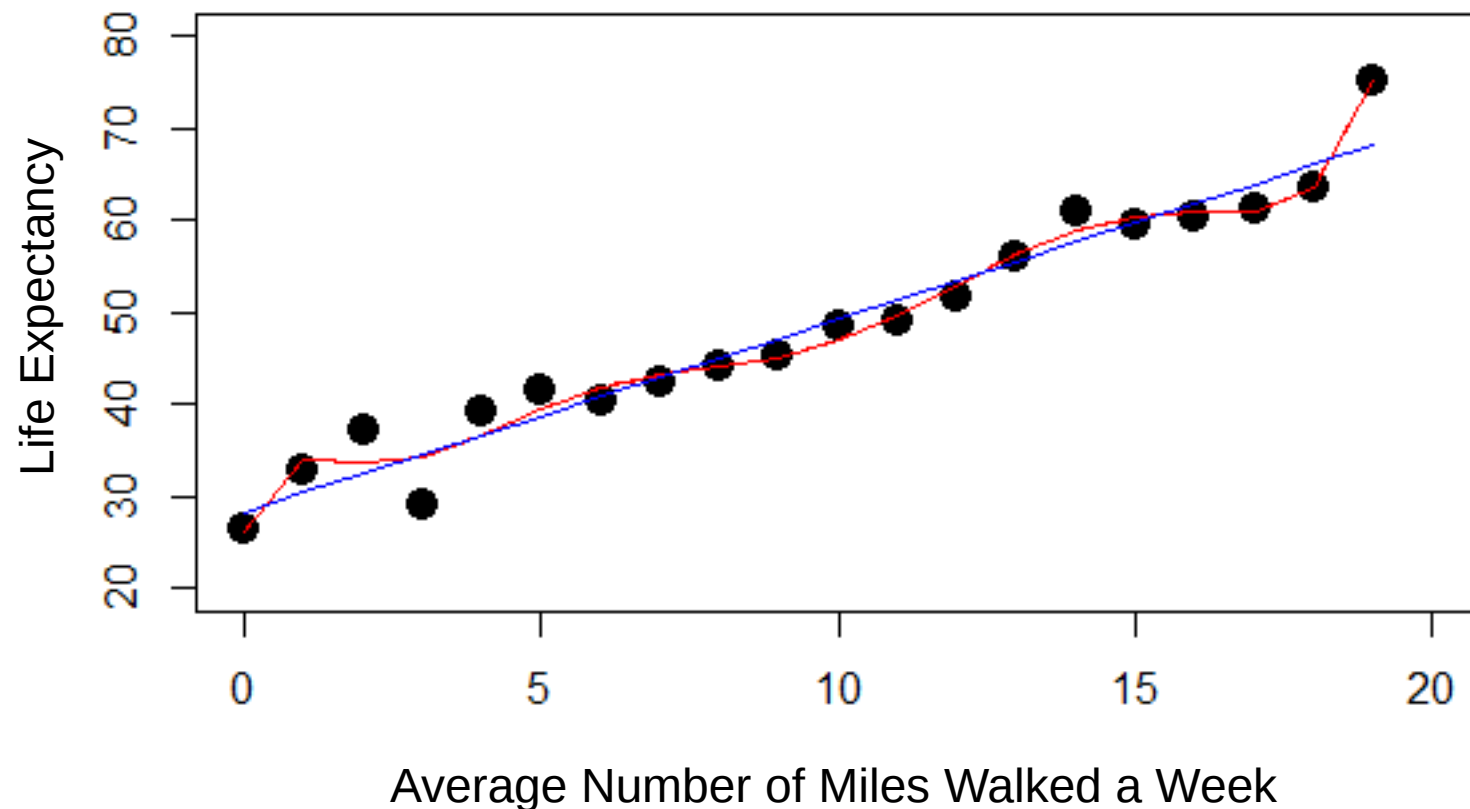


# When Does Overfitting Occur?

Sample Size & Model Complexity

Which model fits **new data** better?

Example: Simulated Data  
N: 20



Simple Model  
 $Y = \beta_0 + \beta_1 X$

Complex Model  
 $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_8 X^8$

Simple Model Error  
MSE: 8.45

Complex Model Error  
MSE: 3.27

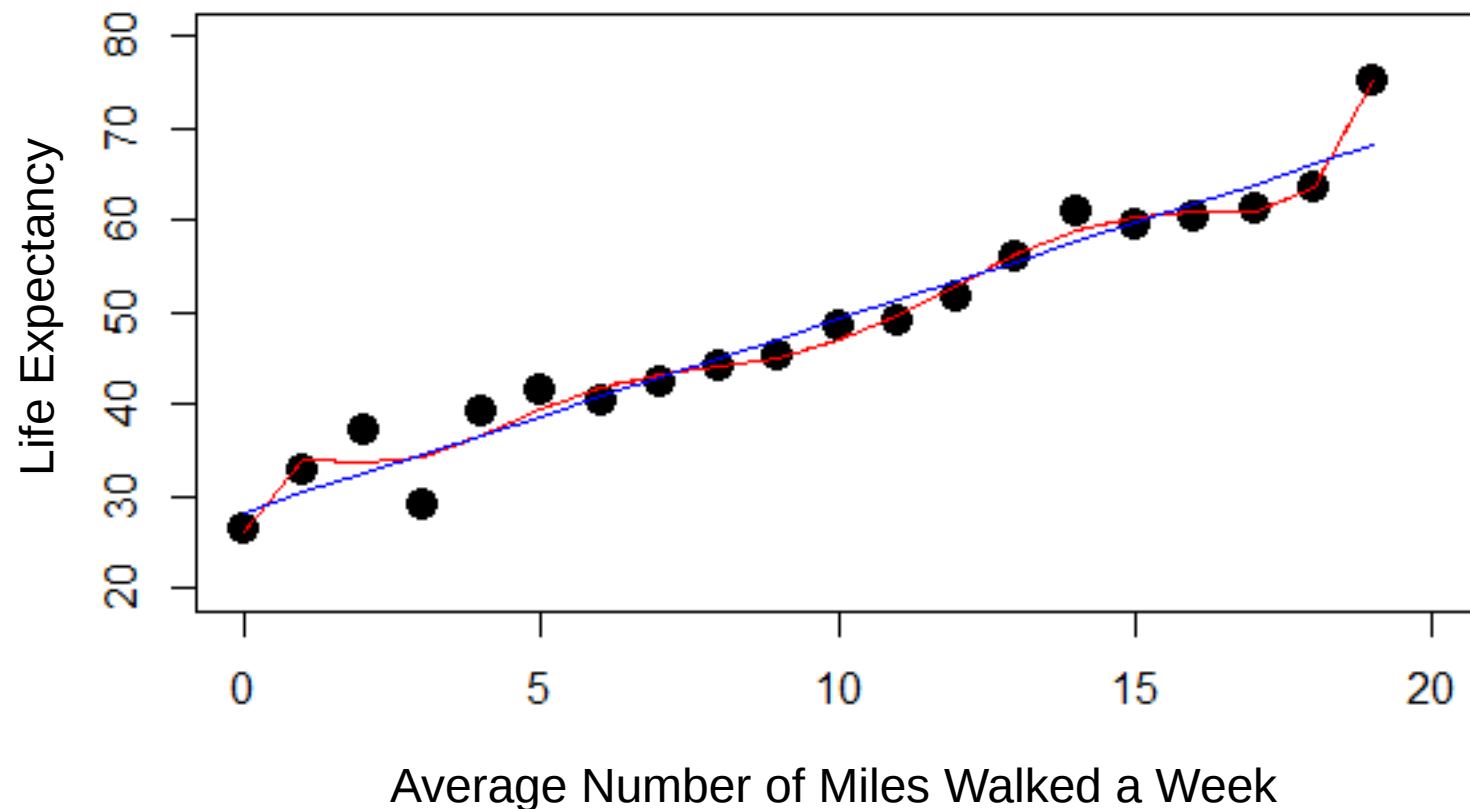


# When Does Overfitting Occur?

Sample Size & Model Complexity

Which model fits **new data** better?

Example: Simulated Data  
N: 20



Simple Model  
 $Y = \beta_0 + \beta_1 X$

Complex Model  
 $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_8 X^8$

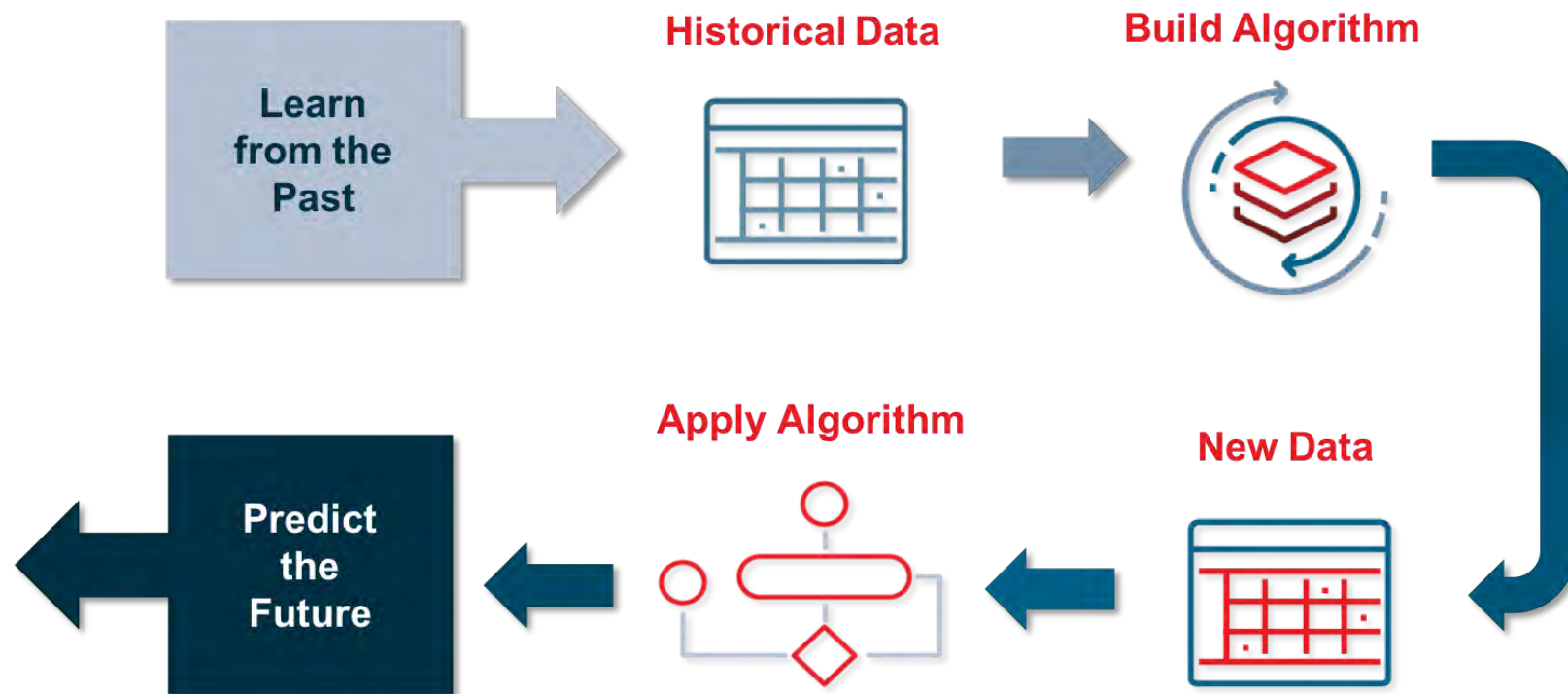


Simple Model Error  
MSE: 8.45

Complex Model Error  
MSE: 3.27

# When Does Overfitting Occur?

Sample Size & Model Complexity

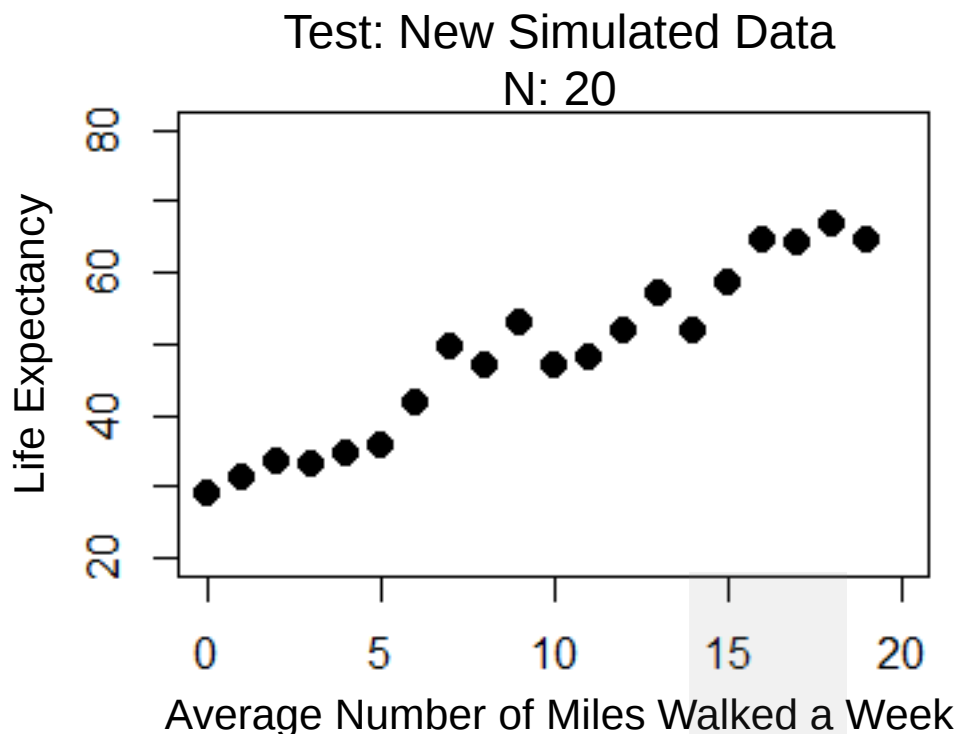
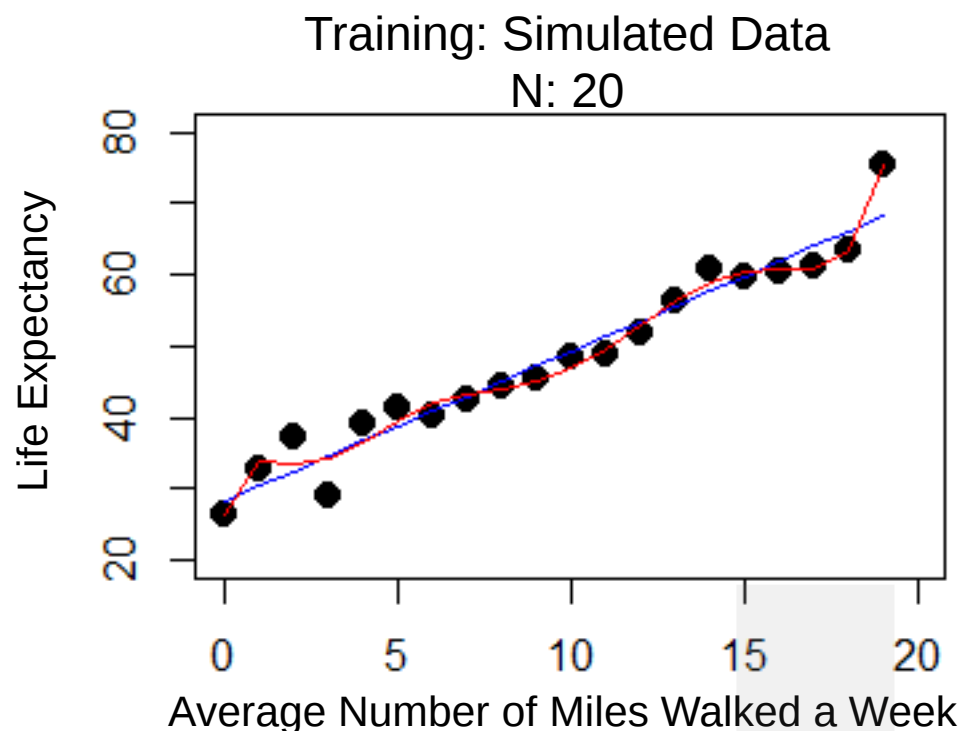


We want to know which model gets us closer to learning about future outcomes and not just our historical data.

Measuring the performance of our models on new data will help us get there.

# When Does Overfitting Occur?

Sample Size & Model Complexity



**Simple Model:**  $Y = \beta_0 + \beta_1 X$

Training MSE: 8.45

Test MSE:

**Complex Model:**

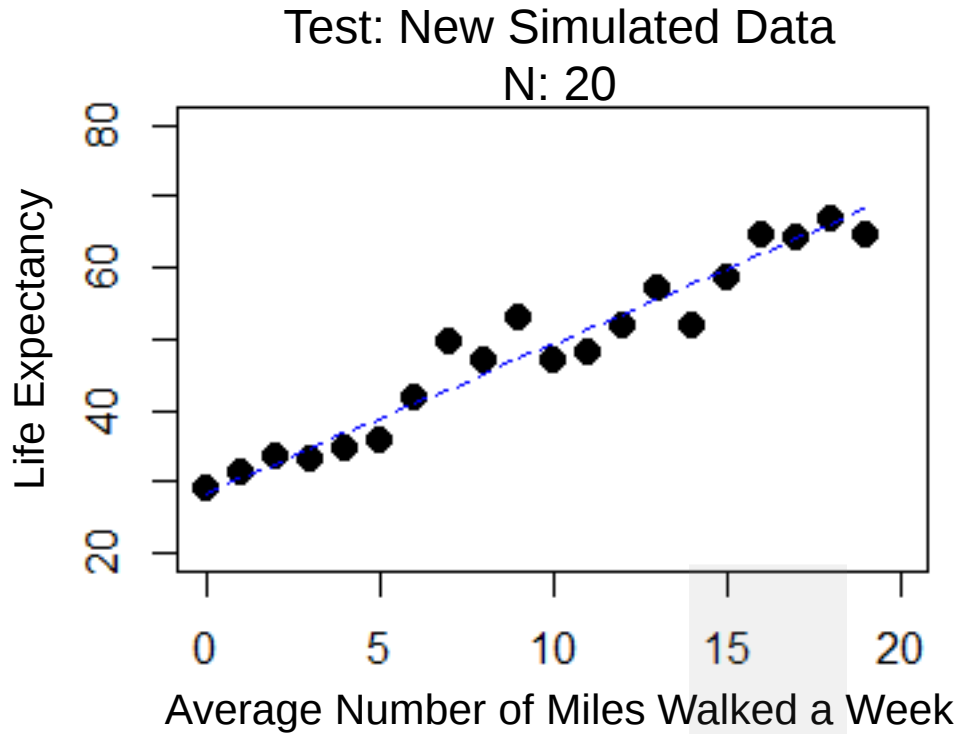
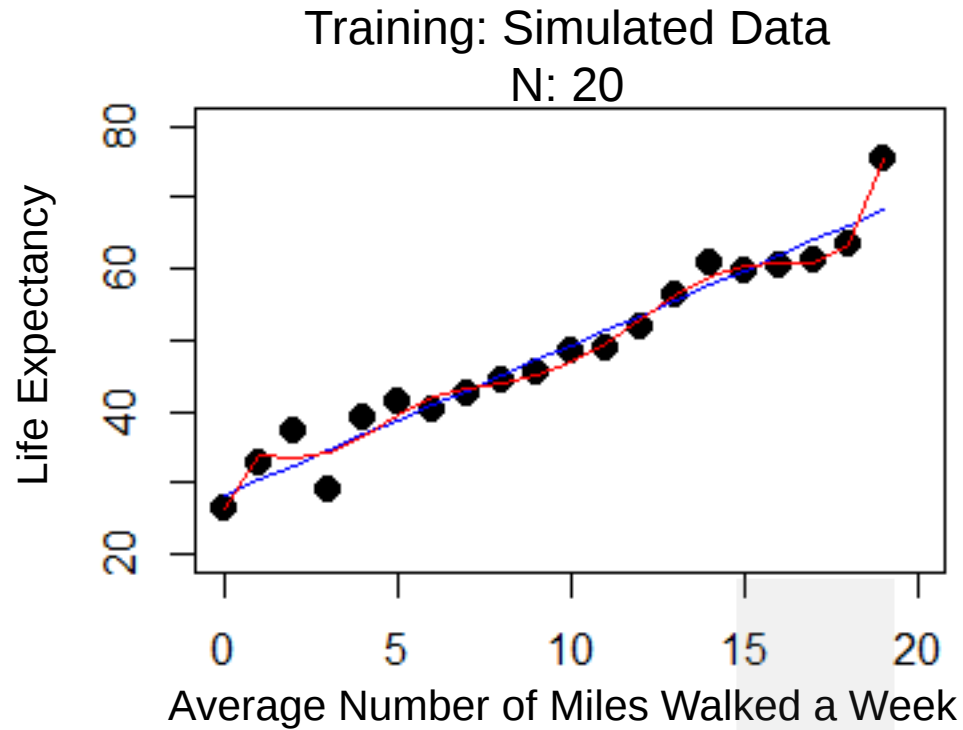
$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_8 X^8$

Training MSE: 3.27

Test MSE:

# When Does Overfitting Occur?

Sample Size & Model Complexity



**Simple Model:**  $Y = \beta_0 + \beta_1 X$

Training MSE: 8.45

Test MSE: 8.86

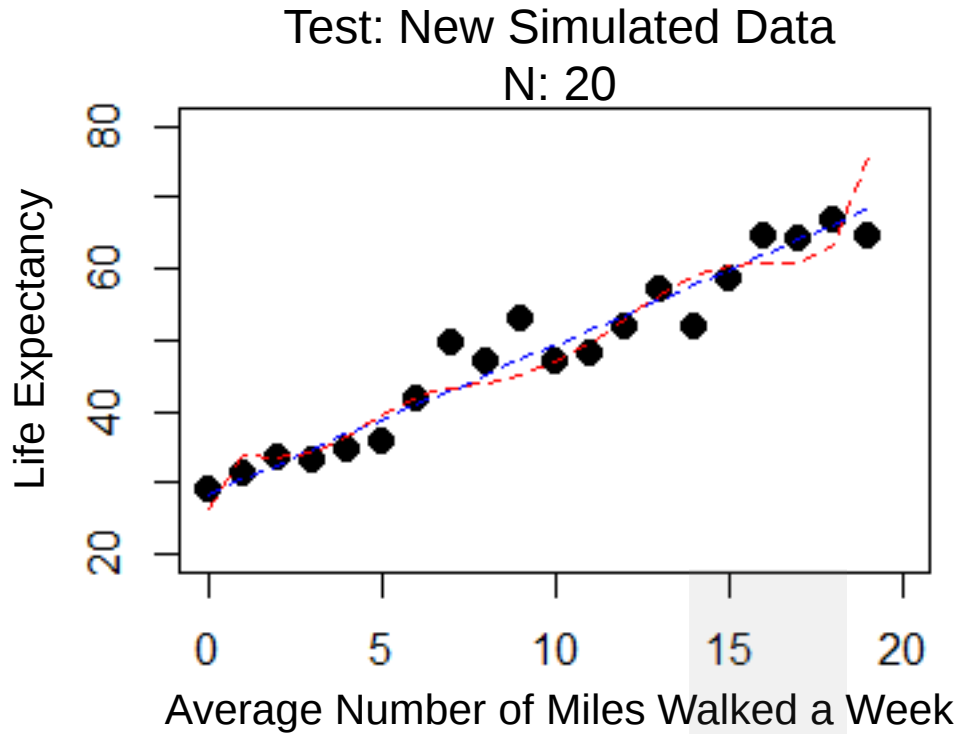
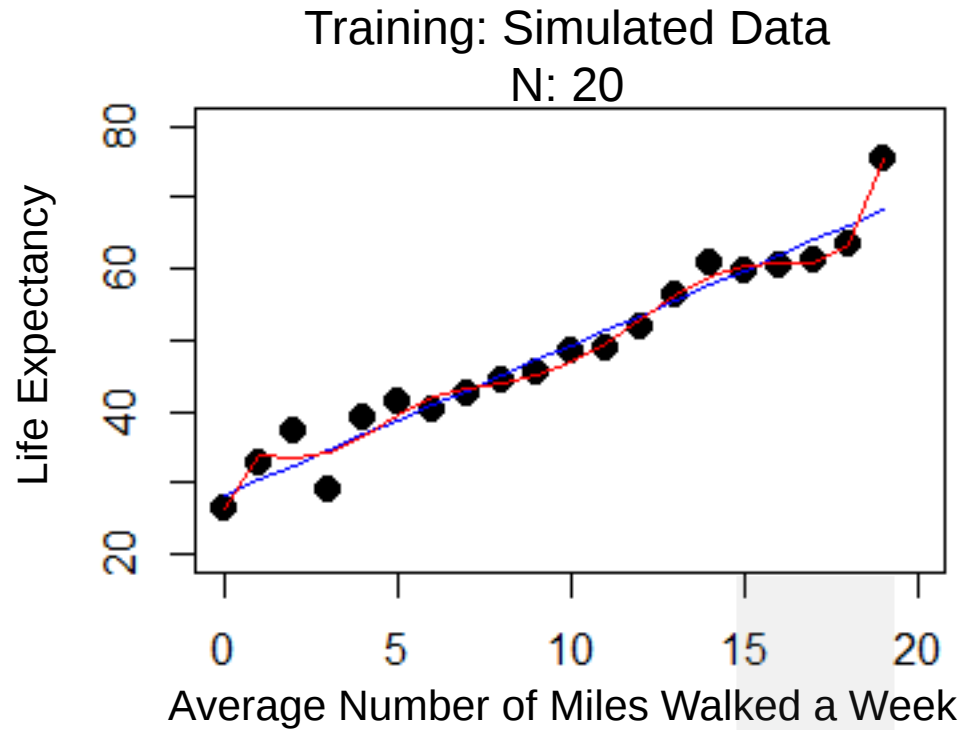
**Complex Model:**  
 $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_8 X^8$

Training MSE: 3.27

Test MSE:

# When Does Overfitting Occur?

Sample Size & Model Complexity



**Simple Model:**  $Y = \beta_0 + \beta_1 X$

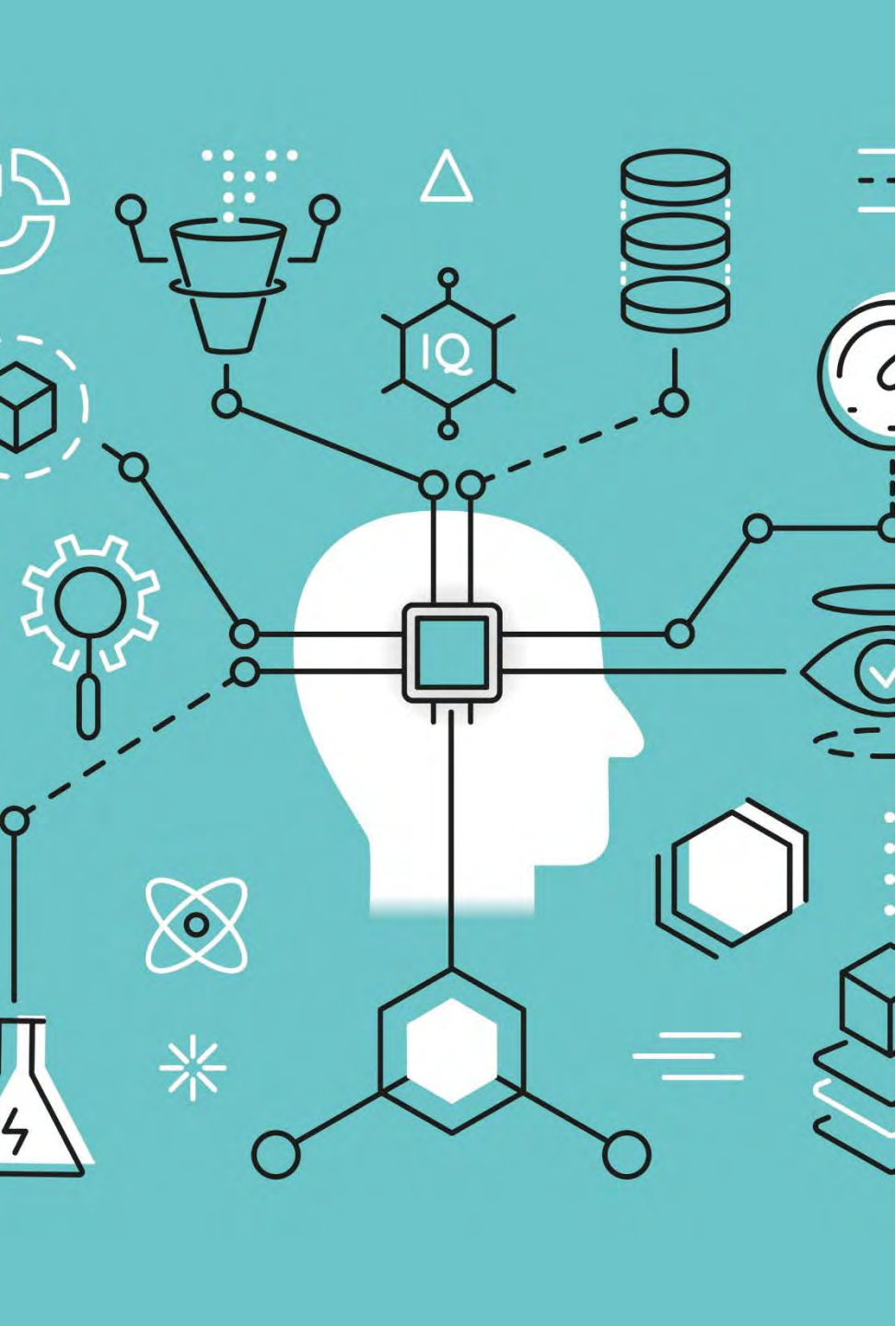
Training MSE: 8.45

✓ Test MSE: 8.86

**Complex Model:**  
 $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_8 X^8$

Training MSE: 3.27

Test MSE: 17.76



**RGA**

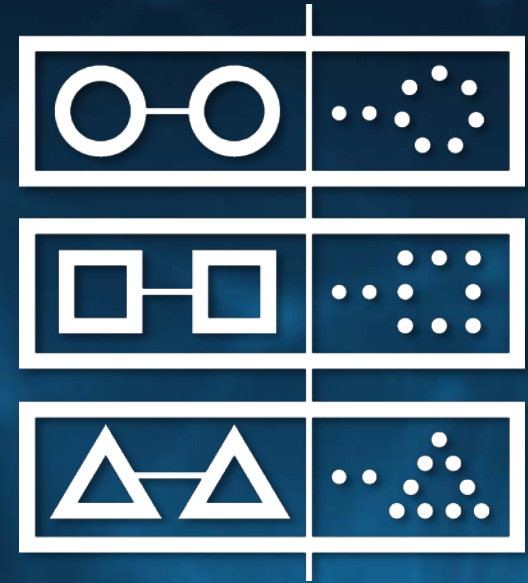
How Do You Prevent Overfitting?

# How Do You Prevent Overfitting?

Testing the procedure on the data that gave it birth is almost certain to overestimate performance.

-Mosteller and Tukey, 1977

If the quantity we care about is how *well* our models will perform on **NEW** data...why don't we just estimate that?



# How Do You Prevent Overfitting?

1

Test set method

2

Leave-one-out Cross Validation

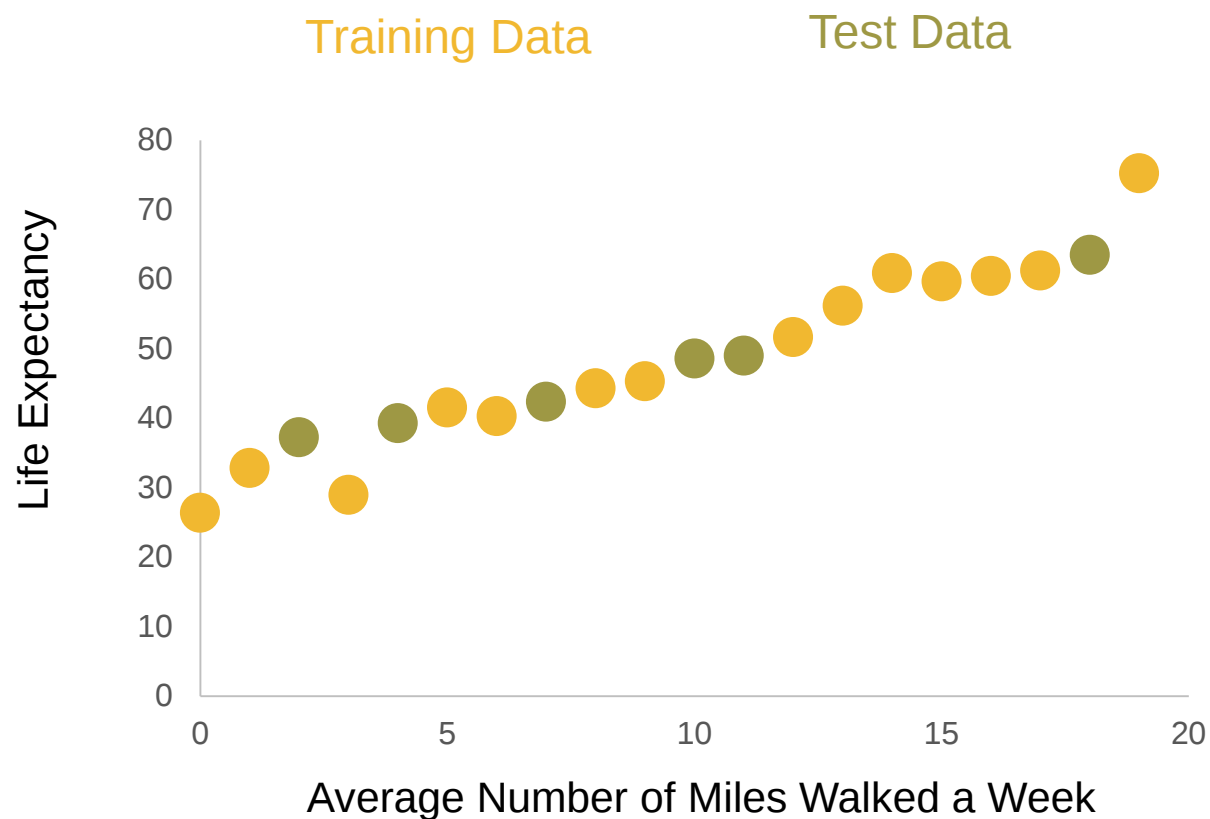
3

K-Fold Cross Validation

Three ways  
to validate  
predictive  
models to  
minimize  
overfitting

# How Do You Prevent Overfitting?

## Test Set Method



1. Randomly select 30% of your data to be your test set
2. Build models on **training data**
3. Estimate future performance by estimating models on **test data**

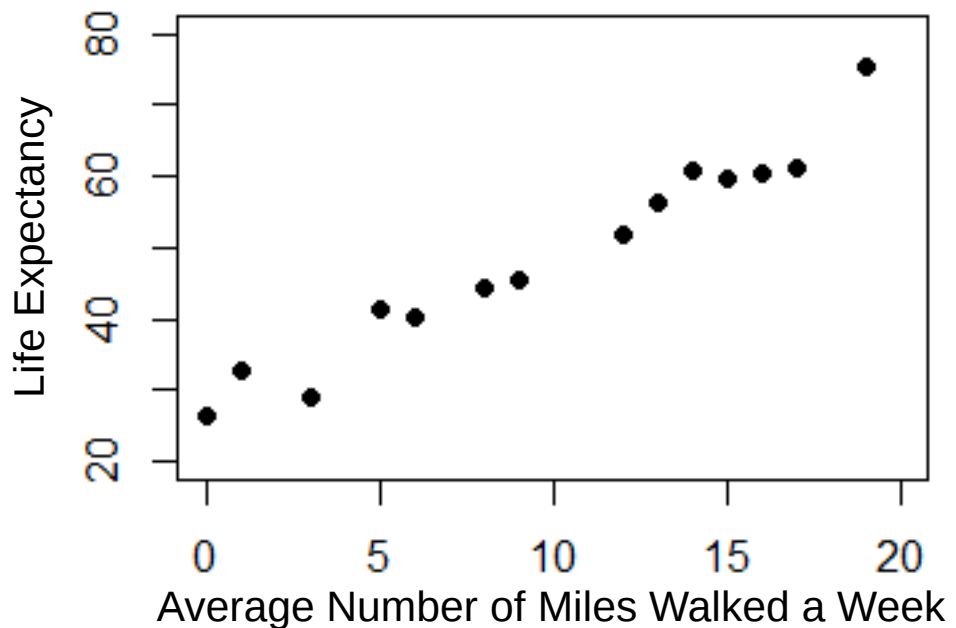
# How Do You Prevent Overfitting?

Test Set Method

1

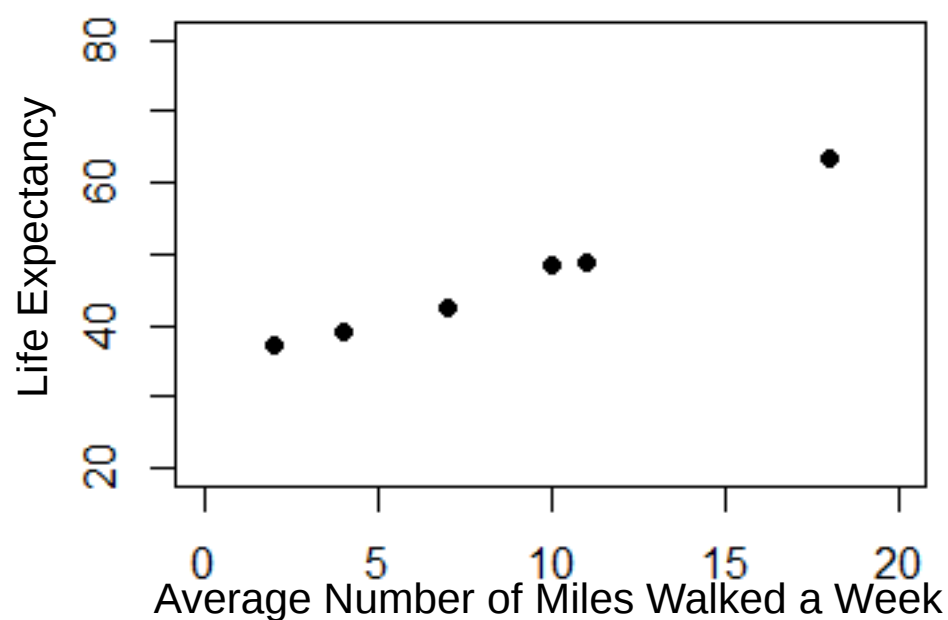
Training Data

N = 14



Test Data

N = 6



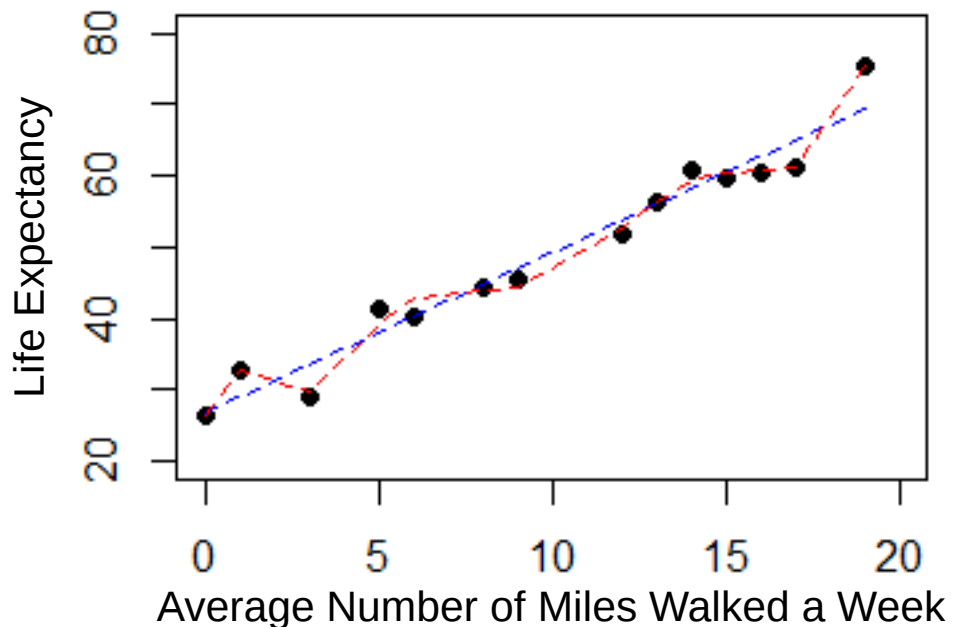
# How Do You Prevent Overfitting?

Test Set Method

1

Training Data

N = 14

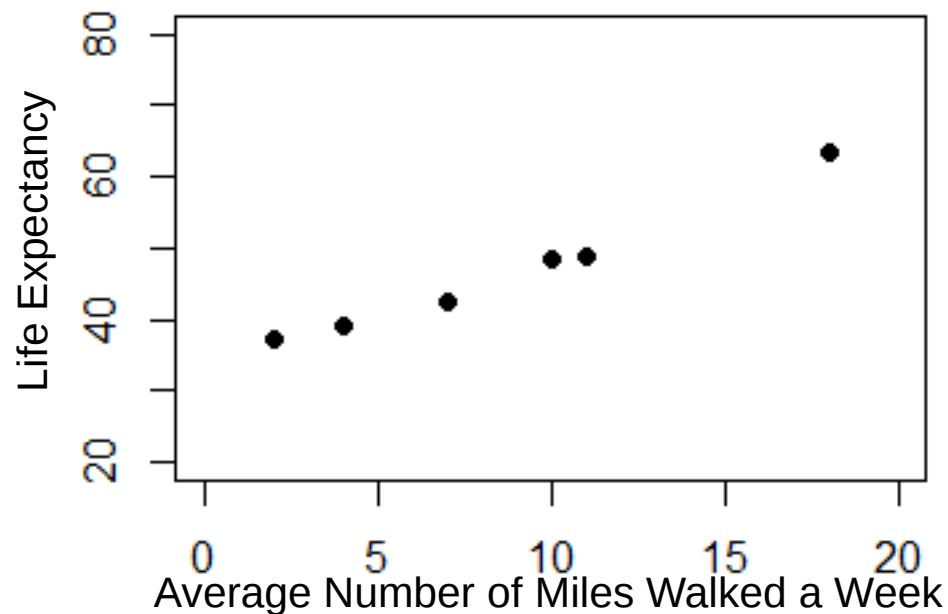


Simple Model Training MSE: 8.21

Complex Model Training MSE: 1.24

Test Data

N = 6



2

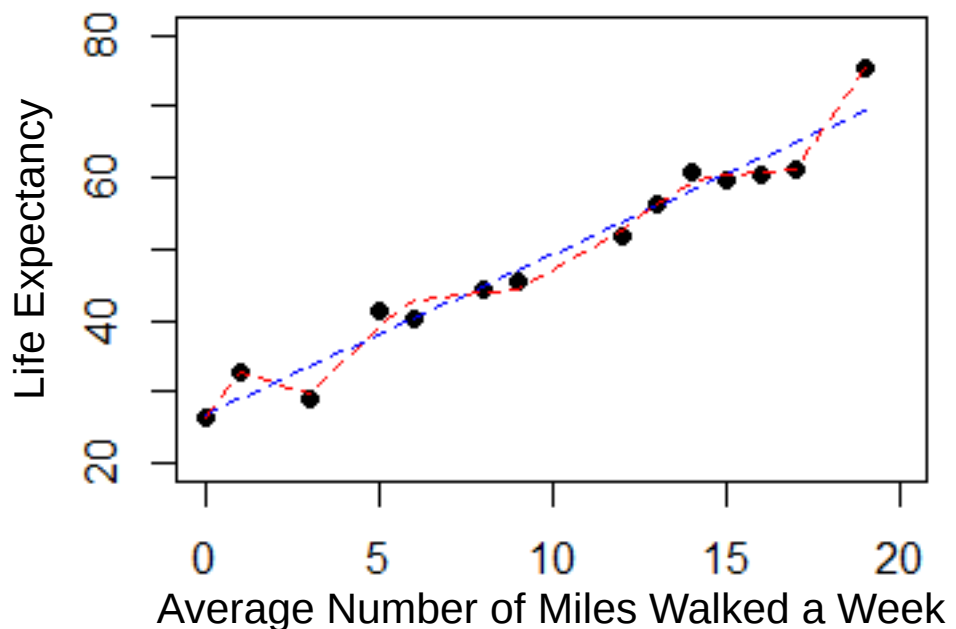
# How Do You Prevent Overfitting?

Test Set Method

1

Training Data

N = 14



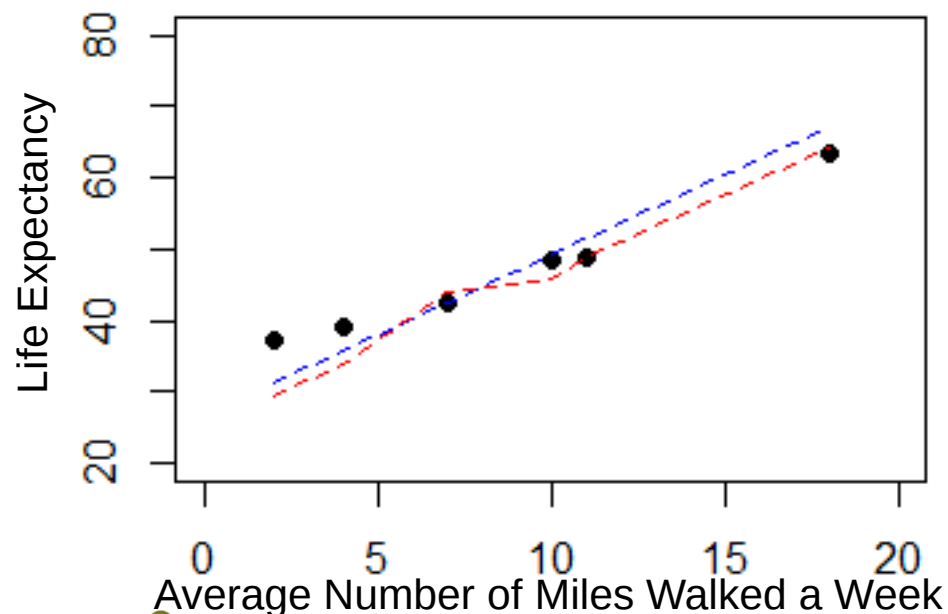
Simple Model Training MSE: 8.21

Complex Model Training MSE: 1.24

2

Test Data

N = 6



Simple Model Test MSE: 11.60

Complex Model Test MSE: 16.95

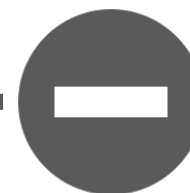
3

# How Do You Prevent Overfitting?

## Test Set Method



Easy to implement



The more data you use to estimate test error, the less data you have to build your model

More data used for training results in more uncertainty about the test error estimate

Less data used for training results in more uncertainty about the model

# How Do You Prevent Overfitting?

1

Test set method

2

Leave-one-out Cross Validation

3

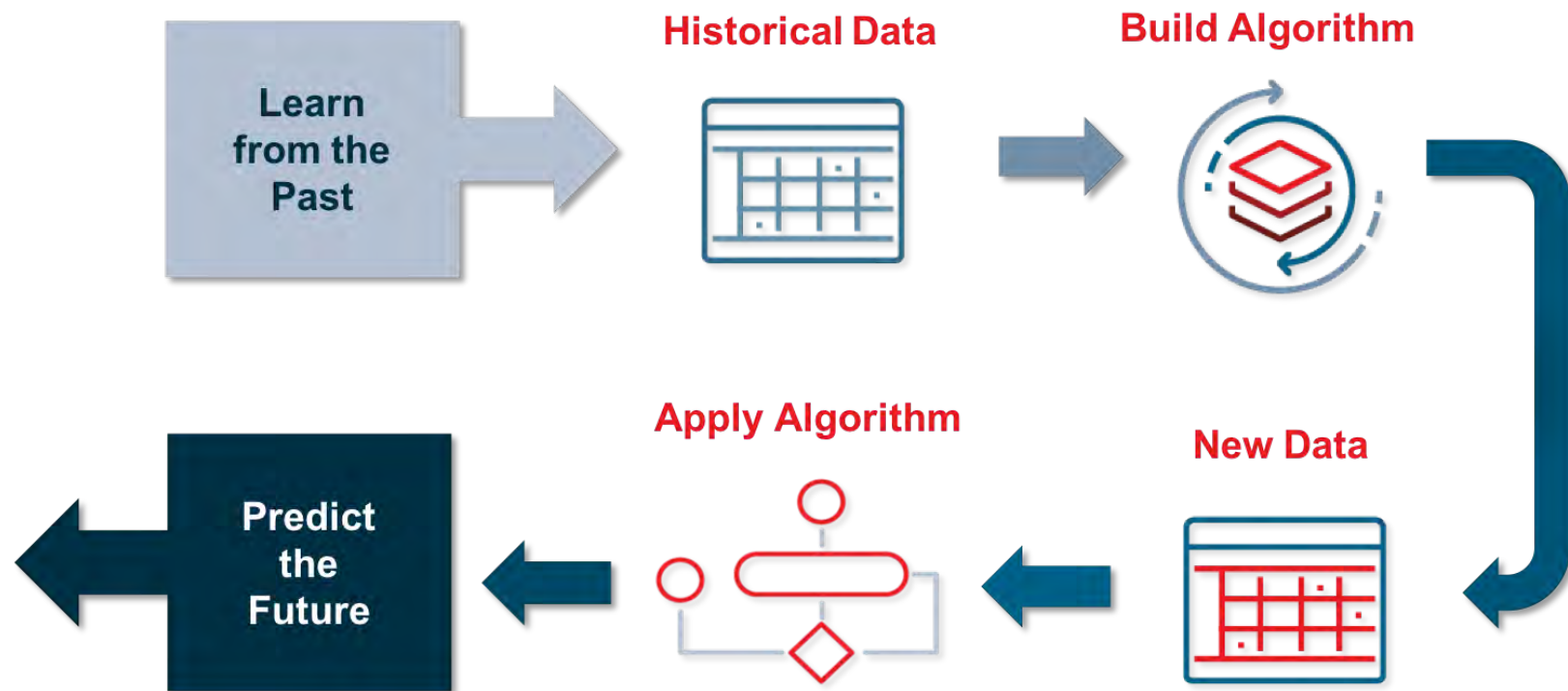
K-Fold Cross Validation

Three ways  
to validate  
predictive  
models to  
minimize  
overfitting

# How Do You Prevent Overfitting?

These are some additional classical ways to approach overfitting and researcher degrees of freedom:

- AIC/BIC metrics
- Bootstrapping
- Bonferroni correction (adjusts for multiple comparisons)



**RGIA**